(12) INTERNATIONAL APPLICATION PUBLISHED UNDER THE PATENT COOPERATION TREATY (PCT)

(71) Applicant (for all designated States except US): DI-
VERSA CORPORATION [US/US]; 4955 Directors
Place, San Diego, CA 92121-1609 (US).

(72) Inventors; and
(75) Inventors/Applicants (for US only): STEGE, Justin
[US/US]; 6931 Worchester Place, San Diego, CA 92126
(US). PRESTON, Lori, [US/US]; 7655 Palmilla Drive
#4304, San Diego, CA 92122 (US). WEINER, David [/];
13416 Portofino Drive, Del Mar, CA 92014 (US).

(74) Agent: EINHORN, Gregory, P.; Fish & Richardson P.C.,
Suite 500, 4350 La Jolla Village Drive, San Diego, CA
92122 (US).

(54) Title: COMPOSITIONS AND METHODS FOR NORMALIZING ASSAYS

(57) Abstract: In one aspect, the invention provides compositions and methods for expression normalization of enzyme activity
screens. Also provided are methods for method for normalizing gene expression in assays and high throughput enzymatic screens
to measure enzyme activity. Also provided are novel inteins and nucleic acids encoding them.

COMPOSITIONS AND METHODS FOR NORMALIZING ASSAYS

TECHNICAL FIELD

This invention generally relates to molecular genetics and pharmaceutical
chemistry. In one aspect, the invention provides compositions and methods for
expression normalization of enzyme activity screens.

BACKGROUND

Protein splicing involves the removal of an internal protein sequence from
a precursor molecule and the ligation of the two flanking sequences to produce a mature
protein product. This post-translational event is analogous to the removal of an intron
from rRNA. These protein splicing "introns" have been called "inteins." The mechanism
by which protein splicing is achieved may be entirely encoded within the internal protein
sequence, or intein.

Variability of cell growth and gene expression limits the sensitive detection
of changes in enzyme properties. These variations make it difficult to discriminate
between changes in gene expression and changes in specific activity in an enzyme
activity screen. The introduction of expression normalization can improve the sensitivity,
reliability and productivity of any activity screen.

SUMMARY

The invention provides a chimeric protein comprising at least three
domains, wherein the first domain comprises at least one enzyme domain or a binding
protein domain, the second domain comprises at least one intein domain and a third
domain comprising a detectable moiety domain, at least one intein domain is positioned
between at least one enzyme or binding protein and at least one detectable moiety
domain, and the intein domain has at least one cleavage or splicing activity.

In one aspect, the enzyme is a nitrilase, a racemase, a kinase or a
hydrolase, such as epoxide hydrolase, a phosphatase, a lipase or a protease. In one aspect,
the binding protein is an antibody or a receptor.

In one aspect, the intein comprises a polypeptide encoded by a nucleic acid
sequence as set forth in SEQ ID NO:1, or, the intein comprises a sequence as set forth in
SEQ ID NO:2.

1

In one aspect, the detectable moiety domain comprises a detectable peptide or polypeptide. The detectable peptide or a polypeptide can be a fluorescent peptide or polypeptide. The detectable peptide or a polypeptide can be a bioluminescent or a chemiluminescent peptide or polypeptide. In one aspect, the bioluminescent or chemiluminescent polypeptide comprises a green fluorescent protein (GFP), an aequorin, an obelin, a mnemiopsin or a berovin. In one aspect, the detectable moiety domain comprises an enzyme that generates a detectable signal. The enzyme that generates a detectable signal can comprise an alpha-galactosidase, an antibiotic (e.g., chloramphenicol acetyltransferase) or a kinase. The detectable moiety domain can comprise a radioactive isotope.

In one aspect, the chimeric protein is a recombinant fusion protein.

In one aspect, the intein domain splicing activity results in cleavage of the enzyme domain from the intein domain and detectable domain. The intein domain splicing activity can result in cleavage of the enzyme domain from the intein domain and detectable domain and cleavage of the detectable domain from the intein domain. In one aspect, the intein domain splicing activity results in cleavage of the detectable domain from the intein domain. In one aspect, the intein domain has only splicing activity. The intein domain can have only cleaving activity.

In one aspect, at least one domain is separated from another domain by a linker. The linker can be a flexible linker. The intein domain can be separated from the detectable moiety domain and the enzyme domain by a linker.

The invention provides an isolated or recombinant nucleic acid encoding a chimeric protein of the invention, e.g., a chimeric protein comprising at least three domains, wherein the first domain comprises at least one enzyme domain or a binding protein domain, the second domain comprises at least one intein domain and a third domain comprising a detectable moiety domain, at least one intein domain is positioned between at least one enzyme or binding protein and at least one detectable moiety domain, and the intein domain has at least one splicing or cleavage activity.

The invention provides an expression cassette comprising a nucleic acid of the invention, e.g., an isolated or recombinant nucleic acid encoding a chimeric protein comprising at least three domains, wherein the first domain comprises at least one enzyme domain or a binding protein domain, the second domain comprises at least one intein domain and a third domain comprising a detectable moiety domain, at least one intein domain is positioned between at least one enzyme or binding protein and at least

2

one detectable moiety domain, and the intein domain has at least one splicing or cleavage activity.

The invention provides a vector comprising a nucleic acid of the invention, e.g., an isolated or recombinant nucleic acid encoding a chimeric protein comprising at least three domains, wherein the first domain comprises at least one enzyme domain or a binding protein domain, the second domain comprises at least one intein domain and a third domain comprising a detectable moiety domain, at least one intein domain is positioned between at least one enzyme or binding protein and at least one detectable moiety domain, and the intein domain has at least one splicing activity or one cleaving activity.

The invention provides a cell comprising a nucleic acid of the invention, e.g., a nucleic acid encoding a chimeric protein comprising at least three domains, wherein the first domain comprises at least one enzyme domain or a binding protein domain, the second domain comprises at least one intein domain and a third domain comprising a detectable moiety domain, at least one intein domain is positioned between at least one enzyme or binding protein and at least one detectable moiety domain, and the intein domain has at least one splicing or cleavage activity.

The invention provides a non-human transgenic animal comprising a nucleic acid of the invention, e.g., a nucleic acid encoding a chimeric protein comprising at least three domains, wherein the first domain comprises at least one enzyme domain or a binding protein domain, the second domain comprises at least one intein domain and a third domain comprising a detectable moiety domain, at least one intein domain is positioned between at least one enzyme or binding protein and at least one detectable moiety domain, and the intein domain has at least one splicing or cleavage activity.

The invention provides a method for normalizing gene expression comprising the following steps: (a) providing a nucleic acid encoding a chimeric protein comprising at least three domains, wherein the first domain comprises at least one enzyme domain or a binding protein domain, the second domain comprises at least one intein domain and a third domain comprising a detectable moiety domain, at least one intein domain is positioned between at least one enzyme or binding protein and at least one detectable moiety domain, and the intein domain has at least one splicing or cleavage activity; (b) expressing the nucleic acid such that the chimeric protein is expressed and the intein domain has at least one splicing activity; and (c) measuring both the activity of the enzyme or the binding protein and the amount of detectable moiety domain expressed,

thereby normalizing gene expression. In one aspect, the detectable moiety domain comprises a detectable peptide or polypeptide. The detectable peptide or a polypeptide can be a fluorescent peptide or polypeptide. The fluorescent peptide or polypeptide can be expressed in a cell and the fluorescence can be measured by a FACS or equivalent device.

In one aspect, the nucleic acid is expressed *in vivo*, or, the nucleic acid is expressed *in vitro*.

The invention provides a high throughput enzymatic screen to measure enzyme activity comprising the following steps: (a) providing a nucleic acid or expression vehicle (e.g., a vector) of the invention, e.g., encoding a chimeric protein comprising at least three domains, wherein the first domain comprises at least one enzyme domain or a binding protein domain, the second domain comprises at least one intein domain and a third domain comprising a detectable moiety domain, at least one intein domain is positioned between at least one enzyme or binding protein and at least one detectable moiety domain, and the intein domain has at least one splicing or cleavage activity; (b) expressing the nucleic acid or expression vehicle (e.g., a vector) *in vivo* or *in vitro*; and (c) measuring both the activity of the enzyme and the amount of detectable moiety domain expressed.

The invention provides a high throughput binding protein screens to measure binding activity comprising the following steps: (a) providing a nucleic acid or expression vehicle (e.g., a vector) of the invention, e.g., encoding a chimeric protein comprising at least three domains, wherein the first domain comprises at least one enzyme domain or a binding protein domain, the second domain comprises at least one intein domain and a third domain comprising a detectable moiety domain, at least one intein domain is positioned between at least one enzyme or binding protein and at least one detectable moiety domain, and the intein domain has at least one splicing or cleavage activity; (b) expressing the nucleic acid or expression vehicle (e.g., a vector) *in vivo* or *in vitro*; and (c) measuring both the activity of the binding protein and the amount of detectable moiety domain expressed.

The invention provides an isolated or recombinant polypeptide comprising (or consisting of) (a) a polypeptide comprising an amino acid sequence having at least 90% identity to SEQ ID NO:2 over a region of at least about 100 residues, or (b) a polypeptide encoded by a nucleic acid comprising (i) a nucleic acid sequence having at least 90% sequence identity to SEQ ID NO:1 over a region of at least about 100 residues;

or, (ii) a nucleic acid that hybridizes under stringent conditions to a nucleic acid comprising a sequence as set forth in SEQ ID NO:1, or a subsequence thereof, wherein the sequence identities are determined by analysis with a sequence comparison algorithm or by a visual inspection and the polypeptide has an intein activity. In alternative aspects, the intein activity is a cleaving activity or a splicing activity. The polypeptide can comprise (or consisting of) an amino acid sequence having at least 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 98%, 99%, or more sequence identity to SEQ ID NO:2 over a region of at least about 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600, 650, or more residues. The polypeptide can have the amino acid sequence as set forth in SEQ ID NO:2.

        In one aspect, the isolated or recombinant intein polypeptide is not associated with any sequence to which it is naturally associated (e.g., an *Aquifex* sequence) on its amino terminal end, on its carboxy terminal end, or on both ends of the intein. In one aspect, the invention provides an isolated or recombinant polypeptide comprising (or consisting of) a polypeptide comprising an amino acid sequence having at least 90% identity to SEQ ID NO:2 over a region of at least about 100 residues with the proviso that it is not associated with any sequence to which it is naturally associated (e.g., an *Aquifex* sequence) on its amino terminal end, on its carboxy terminal end, or on both ends of the intein.

        The invention provides an isolated or recombinant nucleic acid comprising (or consisting of) a nucleic acid sequence having at least 90% sequence identity to SEQ ID NO:1 over a region of at least about 100 residues, wherein the nucleic acid encodes at least one polypeptide having an intein activity, and the sequence identities are determined by analysis with a sequence comparison algorithm or by a visual inspection. The nucleic acid can comprise a sequence having at least 55%, 60%, 65%, 70%, 75%, 80%, 85%, 90%, 95%, 98%, 99%, or more sequence identity to SEQ ID NO:1 over a region of at least about 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600, 650, 700, 800, 900 or 1000 or more residues. In one aspect, the sequence comparison algorithm is a BLAST version 2.2.2 algorithm where a filtering setting is set to blastall -p blastp -d "nr pataa" -F F, and all other options are set to default.

        In one aspect, the isolated or recombinant nucleic acid is not associated with any sequence to which it is naturally associated (e.g., an *Aquifex* sequence) on its 3' terminal end, on its 5' terminal end, or on both ends of the intein coding sequence. In one aspect, the invention provides an isolated or recombinant nucleic acid comprising (or

consisting of) a sequence having at least 90% identity to SEQ ID NO:1 over a region of at least about 100 residues with the proviso that it is not associated with any sequence to which it is naturally associated (e.g., an *Aquifex* sequence) on its 3' end, on its 5' end, or on both ends of the intein-coding sequence.

The invention provides an isolated or recombinant nucleic acid, wherein the nucleic acid comprises a sequence that hybridizes under stringent conditions to a nucleic acid comprising a sequence as set forth in SEQ ID NO:1. In one aspect, the nucleic acid is at least about 50, 100, 150, 200, 250, 300, 350, 400, 450, 500, 550, 600, 650, 700, 800, 900 or 1000 or more residues in length or the full length of the gene or transcript. The stringent conditions can include a wash step comprising a wash in 0.2X SSC at a temperature of about 65°C for about 15 minutes.

The invention provides a nucleic acid probe for identifying a nucleic acid encoding a polypeptide with an intein activity, wherein the probe comprises at least 10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 150, 200, 250, 300 or more consecutive bases of a sequence comprising a sequence as set forth in SEQ ID NO:1.

The invention provides an amplification primer sequence pair for amplifying a nucleic acid encoding a polypeptide having an intein activity, wherein the primer pair is capable of amplifying a nucleic acid comprising a sequence as set forth in SEQ ID NO:1.

The invention provides a method of amplifying a nucleic acid encoding a polypeptide having an intein activity comprising amplification of a template nucleic acid with an amplification primer sequence pair capable of amplifying a nucleic acid sequence as set forth in SEQ ID NO:1, or a subsequence thereof.

The invention provides an expression cassette comprising a nucleic acid of the invention, e.g., a sequence having at least 90% sequence identity to SEQ ID NO:1 over a region of at least about 100 residues, wherein the nucleic acid encodes at least one polypeptide having an intein activity, and the sequence identities are determined by analysis with a sequence comparison algorithm or by a visual inspection.

The invention provides a vector comprising a nucleic acid sequence of the invention, e.g., a nucleic acid having at least 90% sequence identity to SEQ ID NO:1 over a region of at least about 100 residues, wherein the nucleic acid encodes at least one polypeptide having an intein activity, and the sequence identities are determined by analysis with a sequence comparison algorithm or by a visual inspection.

The invention provides a cloning vehicle comprising a vector of the invention, e.g., as set forth in claim 65, wherein the cloning vehicle comprises a viral vector, a plasmid, a phage, a phagemid, a cosmid, a fosmid, a bacteriophage or an artificial chromosome.

The invention provides a transformed cell comprising an expression cassette of the invention or a vector of the invention. The cell can be a bacterial cell, a mammalian cell, a fungal cell, a yeast cell, an insect cell or a plant cell.

The invention provides a transgenic non-human animal comprising a nucleic acid of the invention, e.g., a sequence having at least 90% sequence identity to SEQ ID NO:1 over a region of at least about 100 residues, wherein the nucleic acid encodes at least one polypeptide having an intein activity, and the sequence identities are determined by analysis with a sequence comparison algorithm or by a visual inspection.

The invention provides a transgenic plant comprising a nucleic acid of the invention, e.g., a sequence having at least 90% sequence identity to SEQ ID NO:1 over a region of at least about 100 residues, wherein the nucleic acid encodes at least one polypeptide having an intein activity, and the sequence identities are determined by analysis with a sequence comparison algorithm or by a visual inspection.

The invention provides a method of inhibiting the translation of an intein message in a cell comprising administering to the cell or expressing in the cell an antisense oligonucleotide comprising a nucleic acid of the invention, e.g., a sequence complementary to or capable of hybridizing under stringent conditions to a nucleic acid having at least 90% sequence identity to SEQ ID NO:1 over a region of at least about 100 residues, wherein the nucleic acid encodes at least one polypeptide having an intein activity, and the sequence identities are determined by analysis with a sequence comparison algorithm or by a visual inspection.

The invention provides an isolated or recombinant antibody that specifically binds to a polypeptide of the invention or to a polypeptide encoded by a nucleic acid of the invention.

The details of one or more embodiments of the invention are set forth in the accompanying drawings and the description below. Other features, objects, and advantages of the invention will be apparent from the description and drawings, and from the claims.

DESCRIPTION OF DRAWINGS

Figure 1, a schematic describing the development of an intein expression vector, as set forth in detail in Example 1, below.

Figure 2 is a diagram of the constructs used to generate exemplary intein expression vectors, as set forth in detail in Example 1, below.

Figure 3 is a schematic diagram of splicing versus cleavage of the *Aquifex aeolicus* intein, as set forth in detail in Example 1, below.

Figure 4 is an illustration whole cell extracts as analyzed by SDS-PAGE, as set forth in detail in Example 1, below.

Figure 5 illustrates the overexpression of VMA intein in BL21(DE3)RIL, as set forth in detail in Example 1, below.

Figure 6 illustrates exemplary chimeric structures that retain intein function and fluorescence, as set forth in detail in Example 1, below. Figure 6A illustrates flexible linkers introduced between the intein and FP domains. Figure B illustrates fusion of the GFP C-terminal to the intein, as set forth in detail in Example 1, below.

Figure 7 illustrates β-gal fusions to VMA-FP inteins, as set forth in detail in Example 1, below.

Figure 8 schematically illustrates an exemplary method of the invention for making libraries of inteins using GeneReassembly™. Figure 8A shows recombination locations between essential splicing motifs. Figure 8B schematically illustrates an exemplary method of making a complex library of all possible combinations of splicing motifs, as set forth in detail in Example 1, below.

Figure 9 illustrates an exemplary method of the invention where at the same time that the intein deletions are screened a variety of linkers can be added to the N- and/or C- terminal of the detectable protein, as set forth in detail in Example 1, below.

Figure 10 schematically illustrates an exemplary method of the invention that creates a library in intein mutants using gene mutation (GSSM) and gene reassembly, antibiotic selection with kanamycin followed by kanamycin resistant activity assays to generate inteins with increased splicing efficiency, as set forth in detail in Example 1, below.

Figure 11 schematically illustrates an exemplary method of the invention that creates a library in intein mutants using gene mutation (GSSM) and gene reassembly, antibiotic selection with kanamycin followed by kanamycin resistant activity assays to generate inteins with increased cleavage efficiency, wherein the intein is fused between

8

the target gene as an "N-extein" and a His6 affinity tag as a "C-extein," as set forth in detail in Example 1, below.

Figure 12 schematically illustrates an exemplary method of the invention to quantitate the enantioselective conversion of amino acids using secondary detection enzymes, as set forth in detail in Example 1, below.

Figure 13 schematically illustrates an exemplary nitrilase activity assay of the invention using the fluorogenic reagent dihydrophenoxazine, as set forth in detail in Example 1, below.

Figure 14 schematically illustrates exemplary assays of the invention: a chromogenic epoxide assay and a fluorogenic epoxide assay, in Figure 14A and Figure 14B, respectively.

Like reference symbols in the various drawings indicate like elements.

DETAILED DESCRIPTION

The invention provides compositions and methods for expression normalization of enzyme activity screens. The invention provides chimeric polypeptides ("fusion proteins") with at least three domains: a first domain comprising at least one enzyme, a second domain between the first and the third domain comprising at least one intein and a third domain comprising at least one detectable moiety (domain), e.g., a protein such as a fluorescent domain or a bioluminescent protein, or any other detectable moiety.

In one aspect, the chimeric polypeptide of the invention is expressed recombinantly *in vitro* or in a cell, i.e., *in vivo*. After expression of the protein the cleavage activity of the intein moiety produces a stoichiometric expression of the two proteins, i.e., the enzyme of the first domain and the detectable moiety of the third domain. Thus, the chimeric polypeptides of the invention can be used for expression normalization of enzyme activity, e.g., in assays and screens. The compositions and methods of the invention can improve the sensitivity, reliability and productivity of any activity screen. The compositions and methods of the invention can be used in high throughput enzymatic screens to measure enzyme activities in whole cells or cell lysates. The compositions and methods of the invention can be used to monitor enzymatic conversions directly or indirectly in a single cell assay, e.g., a FACS-based assay, or in a group of cells, e.g., those grown in a well of a microtiter plate.

9

The compositions and methods of the invention can be used to discriminate high specific activity enzymes, e.g., clones, from high expressing enzymes, e.g., clones, even though different genes (or enzyme coding sequences in vectors) can be expressed at dramatically different levels. The compositions and methods of the invention can be used to decrease background noise generated by liquid transfer errors and differences in growth and expression. Thus, in one aspect, the compositions and methods of the invention are used to normalize enzyme activity to gene expression to improve the sensitivity and reliability of an assay, e.g., an enzyme assay. Because the compositions and methods of the invention can be highly sensitive, in one aspect they are used to develop new assays, e.g., enzyme assays, that otherwise might have been impossible due to variations in gene expression. In one aspect, the improved sensitivity allows detection of small activity increases, e.g., the small activity increases that are generated in each cycle of an iterative enzyme evolution. By themselves, small increases in enzyme activity may not significantly improve the value of an enzyme. However, the synergy generated by combining many small mutations iteratively can result in very dramatic improvements in enzyme activity.

In one aspect of the methods of the invention, the fluorescent measurement of gene expression increases the sensitivity and information generated by standard microtiter plate screens for enzyme activity. The screening methods of the invention can completely cover large enzyme libraries, e.g., for enzyme discovery and evolution. In one aspect, the invention provides sub-microliter and single cell assays to detect enzyme activity for discovery purposes. In one aspect, the invention provides quantitative activity measurements for, e.g., detecting desirable mutants and for enzyme evolution activity assays.

In one aspect, the methods normalize expression (e.g., enzyme expression) to fluorescence. Thus, the compositions and methods of the invention can be used for ultra high throughput enzyme screens. In some aspects, this improves throughput by several orders of magnitude and can enable complete sampling of the enzyme libraries.

The compositions and methods of the invention can be used in any assay, including enzymatic screens and spectroscopic assays. In alternative aspects, the compositions and methods of the invention are used in spectroscopic assays, e.g., in microtiter plates, GigaMatrix™, and FACS. In alternative aspects of the methods of the invention, activity is measured by a variety of biological and analytical methods. All

assays providing a quantitative value for activity are incorporated into the methods of the invention to normalize activity to protein expression.

In one aspect, the methods of the invention are used as a general technique to achieve stoichiometric expression of two proteins without the limitations of a fusion protein. In one aspect, the intein domain of a chimeric polypeptide of the invention (e.g., an intein-fluorescent protein) has autocatalytic splicing activity.

The intein domain of a chimeric polypeptide of the invention (e.g., an intein-fluorescent protein) can function equally well in eukaryotic as well as prokaryotic systems. Thus, the compositions and methods of the invention can be applied to any organism.

The chimeric polypeptides of the invention (e.g., intein-fluorescent proteins) can be used to develop, e.g., "evolve," enzymes. For example, the chimeric polypeptides of the invention can be used to "evolve" enzymes for the biocatalytic production of commercial products, e.g., pharmaceuticals and industrial chemicals and proteins. Any enzyme can be linked to a detectable domain via an intein, e.g., a hydrolase, a racemase, a nitrilase and/or a epoxide hydrolase enzyme. Targets of racemases can be proteogenic or non-proteogenic amino acids.

In one aspect, the intein domain of a chimeric (multi-domain) protein of the invention has more cleavage activity than splicing activity. In alternative aspects, the intein domain has greater than 50%, 60%, 70%, 80%, 90%, 95% or 98% or more *in vivo* or *in vitro* splicing activity.

In one aspect of the methods of the invention, a target protein (e.g., a substrate) is used in a purified form. In one aspect, high expressing genes identified.

The invention provides nucleic acids encoding the chimeric polypeptides of the invention. Also provided are expression systems, e.g., expression cassettes, recombinant viruses, vectors, and the like for expressing and using the chimeric proteins of the invention in enzyme activity assays. The coding sequences can be operably linked to a transcriptional control sequence, such as a promoter, e.g., a constitutive or an inducible promoter.

Definitions

The term "intein" includes all polypeptides capable of any intein activity, e.g., capable of protein splicing or protein cleaving involving the removal of an internal

protein sequence. In one aspect, the intein is also capable of ligating the two flanking
sequences.

        The invention provides expression cassettes comprising nucleic acids of
the invention. The term "expression cassette" includes nucleotide sequences which are
5   capable of affecting expression of a coding sequence (e.g., a chimeric polypeptide of the
invention) in a host compatible with such sequences. Expression cassettes include at least
a promoter operably linked with the polypeptide coding sequence; and, optionally, with
other sequences, e.g., transcription termination signals. Additional factors necessary or
helpful in effecting expression may also be used, e.g., enhancers. Thus, expression
10   cassettes also include plasmids, expression vectors, recombinant viruses, any form of
recombinant "naked DNA" vector, and the like. A "vector" can comprise a nucleic acid
which can infect, transfect, transiently or permanently transduce a cell. It will be
recognized that a vector can be a naked nucleic acid, or a nucleic acid complexed with
protein or lipid. The vector optionally comprises viral or bacterial nucleic acids and/or
15   proteins, and/or membranes (e.g., a cell membrane, a viral lipid envelope, etc.). Vectors
include, but are not limited to replicons (e.g., RNA replicons, bacteriophages) to which
fragments of DNA may be attached and become replicated. Vectors thus include, but are
not limited to RNA, autonomous self-replicating circular or linear DNA or RNA (e.g.,
plasmids, viruses, and the like, see, e.g., U.S. Patent No. 5,217,879), and include both the
20   expression and non-expression plasmids. Where a recombinant microorganism or cell
culture is described as hosting an "expression vector" this includes both extra-
chromosomal circular and linear DNA and DNA that has been incorporated into the host
chromosome(s). Where a vector is being maintained by a host cell, the vector may either
be stably replicated by the cells during mitosis as an autonomous structure, or is
25   incorporated within the host's genome.

        The invention provides nucleic acids operably linked to a coding sequence
of the invention. "Operably linked" as used herein can refer to a functional relationship
between two or more nucleic acid (e.g., DNA) segments. Typically, it refers to the
functional relationship of transcriptional regulatory sequence to a transcribed sequence.
30   For example, a promoter is operably linked to a coding sequence, such as a nucleic acid
of the invention, if it stimulates or modulates the transcription of the coding sequence in
an appropriate host cell or other expression system. Generally, promoter transcriptional
regulatory sequences that are operably linked to a transcribed sequence are physically
contiguous to the transcribed sequence, i.e., they are cis-acting. However, some

transcriptional regulatory sequences, such as enhancers, need not be physically contiguous or located in close proximity to the coding sequences whose transcription they enhance. As used herein, the term "promoter" can include all sequences capable of driving transcription of a coding sequence in a cell, e.g., a plant cell. Thus, promoters used in the constructs of the invention include *cis*-acting transcriptional control elements and regulatory sequences that are involved in regulating or modulating the timing and/or rate of transcription of a gene. For example, a promoter can be a *cis*-acting transcriptional control element, including an enhancer, a promoter, a transcription terminator, an origin of replication, a chromosomal integration sequence, 5' and 3' untranslated regions, or an intronic sequence, which are involved in transcriptional regulation. These *cis*-acting sequences typically interact with proteins or other biomolecules to carry out (turn on/off, regulate, modulate, etc.) transcription. "Constitutive" promoters are those that drive expression continuously under most environmental conditions and states of development or cell differentiation. "Inducible" or "regulatable" promoters direct expression of the nucleic acid of the invention under the influence of environmental conditions or developmental conditions. Examples of environmental conditions that may affect transcription by inducible promoters include anaerobic conditions, elevated temperature, drought, or the presence of light.

The phrases "nucleic acid" or "nucleic acid sequence" can include oligonucleotide, nucleotide, polynucleotide, or to a fragment of any of these, to DNA or RNA (e.g., mRNA, rRNA, tRNA) of genomic or synthetic origin which may be single-stranded or double-stranded and may represent a sense or antisense strand, to peptide nucleic acid (PNA), or to any DNA-like or RNA-like material, natural or synthetic in origin, including, e.g., iRNA, ribonucleoproteins (e.g., iRNPs). The term encompasses nucleic acids, i.e., oligonucleotides, containing known analogues of natural nucleotides. The term also encompasses nucleic-acid-like structures with synthetic backbones, see e.g., Mata (1997) Toxicol. Appl. Pharmacol. 144:189-197; Strauss-Soukup (1997) Biochem. 36:8692-8698; Samstag (1996) Antisense Nucleic Acid Drug Dev 6:153-156.

The term "isolated" can include a material removed from its original environment, e.g., the natural environment if it is naturally occurring. For example, a naturally occurring polynucleotide or polypeptide present in a living animal is not isolated, but the same polynucleotide or polypeptide, separated from some or all of the coexisting materials in the natural system, is isolated. Such polynucleotides could be part of a vector and/or such polynucleotides or polypeptides could be part of a composition,

and still be isolated in that such vector or composition is not part of its natural environment. As used herein, an isolated material or composition can also be a "purified" composition, i.e., it does not require absolute purity; rather, it is intended as a relative definition. Individual nucleic acids obtained from a library can be conventionally purified to electrophoretic homogeneity. In alternative aspects, the invention provides nucleic acids which have been purified from genomic DNA or from other sequences in a library or other environment by at least one, two, three, four, five or more orders of magnitude.

As used herein, the term "recombinant" can include nucleic acids adjacent to a "backbone" nucleic acid to which it is not adjacent in its natural environment. In one aspect, nucleic acids represent 5% or more of the number of nucleic acid inserts in a population of nucleic acid "backbone molecules." "Backbone molecules" according to the invention include nucleic acids such as expression vectors, self-replicating nucleic acids, viruses, integrating nucleic acids, and other vectors or nucleic acids used to maintain or manipulate a nucleic acid insert of interest. In one aspect, the enriched nucleic acids represent 10%, 15%, 20%, 30%, 40%, 50%, 60%, 70%, 80%, 90%, 95%, 98% or more of the number of nucleic acid inserts in the population of recombinant backbone molecules. "Recombinant" polypeptides or proteins refer to polypeptides or proteins produced by recombinant DNA techniques; e.g., produced from cells transformed by an exogenous DNA construct encoding the desired polypeptide or protein. "Synthetic" polypeptides or protein are those prepared by chemical synthesis, as described in further detail, below.

"Amino acid" or "amino acid sequence" can include an oligopeptide, peptide, polypeptide, or protein sequence, or to a fragment, portion, or subunit of any of these, and to naturally occurring or synthetic molecules. The terms "polypeptide" and "protein" include amino acids joined to each other by peptide bonds or modified peptide bonds, i.e., peptide isosteres, and may contain modified amino acids other than the 20 gene-encoded amino acids. The term "polypeptide" also includes peptides and polypeptide fragments, motifs and the like. The term also includes glycosylated polypeptides. The peptides and polypeptides of the invention also include all "mimetic" and "peptidomimetic" forms, as described in further detail, below.

Generating and Manipulating Nucleic Acids

           The invention provides nucleic acids, including expression cassettes such as expression vectors, encoding the chimeric polypeptides of the invention. The invention also includes methods for discovering new chimeric and new intein sequences
5   using the nucleic acids of the invention. Also provided are methods for modifying the nucleic acids of the invention, including the intein of the invention, by, e.g., synthetic ligation reassembly, optimized directed evolution system and/or saturation mutagenesis.

           The nucleic acids of the invention can be made, isolated and/or manipulated by, e.g., cloning and expression of cDNA libraries, amplification of message
10   or genomic DNA by PCR, and the like. In practicing the methods of the invention, homologous genes can be modified by manipulating a template nucleic acid, as described herein. The invention can be practiced in conjunction with any method or protocol or device known in the art, which are well described in the scientific and patent literature.

       General Techniques

15   The nucleic acids used to practice this invention, whether RNA, iRNA, antisense nucleic acid, cDNA, genomic DNA, vectors, viruses or hybrids thereof, may be isolated from a variety of sources, genetically engineered, amplified, and/or expressed/ generated recombinantly. Recombinant polypeptides generated from these nucleic acids can be individually isolated or cloned and tested for a desired activity. Any recombinant
20   expression system can be used, including bacterial, mammalian, yeast, insect or plant cell expression systems.

           Alternatively, these nucleic acids can be synthesized *in vitro* by well-known chemical synthesis techniques, as described in, e.g., Adams (1983) J. Am. Chem. Soc. 105:661; Belousov (1997) Nucleic Acids Res. 25:3440-3444; Frenkel (1995) Free
25   Radic. Biol. Med. 19:373-380; Blommers (1994) Biochemistry 33:7886-7896; Narang (1979) Meth. Enzymol. 68:90; Brown (1979) Meth. Enzymol. 68:109; Beaucage (1981) Tetra. Lett. 22:1859; U.S. Patent No. 4,458,066.

           Techniques for the manipulation of nucleic acids, such as, e.g., subcloning, labeling probes (e.g., random-primer labeling using Klenow polymerase, nick translation,
30   amplification), sequencing, hybridization and the like are well described in the scientific and patent literature, see, e.g., Sambrook, ed., MOLECULAR CLONING: A LABORATORY MANUAL (2ND ED.), Vols. 1-3, Cold Spring Harbor Laboratory, (1989); CURRENT PROTOCOLS IN MOLECULAR BIOLOGY, Ausubel, ed. John

Wiley & Sons, Inc., New York (1997); LABORATORY TECHNIQUES IN
BIOCHEMISTRY AND MOLECULAR BIOLOGY: HYBRIDIZATION WITH
NUCLEIC ACID PROBES, Part I. Theory and Nucleic Acid Preparation, Tijssen, ed.
Elsevier, N.Y. (1993).

Another useful means of obtaining and manipulating nucleic acids used to
practice the methods of the invention, e.g., enzymes, binding proteins and/or inteins, is to
clone from genomic samples, and, if desired, screen and re-clone inserts isolated or
amplified from, e.g., genomic clones or cDNA clones. Sources of nucleic acid used in the
methods of the invention include genomic or cDNA libraries contained in, e.g.,
mammalian artificial chromosomes (MACs), see, e.g., U.S. Patent Nos. 5,721,118;
6,025,155; human artificial chromosomes, see, e.g., Rosenfeld (1997) Nat. Genet. 15:333-
335; yeast artificial chromosomes (YAC); bacterial artificial chromosomes (BAC); P1
artificial chromosomes, see, e.g., Woon (1998) Genomics 50:306-316; P1-derived vectors
(PACs), see, e.g., Kern (1997) Biotechniques 23:120-124; cosmids, recombinant viruses,
phages or plasmids.

In one aspect, a nucleic acid encoding a polypeptide of the invention is
assembled in appropriate phase with a leader sequence capable of directing secretion of
the translated polypeptide or fragment thereof.

The invention provides fusion proteins and nucleic acids encoding them.
A chimeric polypeptide of the invention can be further fused to a heterologous peptide or
polypeptide, such as N-terminal identification peptides which impart desired
characteristics, such as increased stability or simplified purification. The chimeric
polypeptides of the invention can also be synthesized and expressed as fusion proteins
with one or more additional domains linked thereto for, e.g., to more readily isolate a
recombinantly synthesized peptide, to identify and isolate antibodies and antibody-
expressing B cells, and the like. Detection and purification facilitating domains include,
e.g., metal chelating peptides such as polyhistidine tracts and histidine-tryptophan
modules that allow purification on immobilized metals, protein A domains that allow
purification on immobilized immunoglobulin, and the domain utilized in the FLAGS
extension/affinity purification system (Immunex Corp, Seattle WA). The inclusion of a
cleavable linker sequences such as Factor Xa or enterokinase (Invitrogen, San Diego CA)
between a purification domain and the motif-comprising peptide or polypeptide to
facilitate purification. For example, an expression vector can include an epitope-
encoding nucleic acid sequence linked to six histidine residues followed by a thioredoxin

16

and an enterokinase cleavage site (see e.g., Williams (1995) Biochemistry 34:1787-1797; Dobeli (1998) Protein Expr. Purif. 12:404-414). The histidine residues facilitate detection and purification while the enterokinase cleavage site provides a means for purifying the epitope from the remainder of the fusion protein. Technology pertaining to vectors encoding fusion proteins and application of fusion proteins are well described in the scientific and patent literature, see e.g., Kroll (1993) DNA Cell. Biol., 12:441-53.

Transcriptional and translational control sequences

The invention provides nucleic acid (e.g., DNA) sequences of the invention operatively linked to expression (e.g., transcriptional or translational) control sequence(s), e.g., promoters or enhancers, to direct or modulate RNA synthesis/ expression. The expression control sequence can be in an expression vector. Exemplary bacterial promoters include lacI, lacZ, T3, T7, gpt, lambda PR, PL and trp. Exemplary eukaryotic promoters include CMV immediate early, HSV thymidine kinase, early and late SV40, LTRs from retrovirus, and mouse metallothionein I.

Promoters suitable for expressing a polypeptide in bacteria include, e.g., the *E. coli* lac or trp promoters, the lacI promoter, the lacZ promoter, the T3 promoter, the T7 promoter, the gpt promoter, the lambda PR promoter, the lambda PL promoter, promoters from operons encoding glycolytic enzymes such as 3-phosphoglycerate kinase (PGK), and the acid phosphatase promoter. Eukaryotic promoters include the CMV immediate early promoter, the HSV thymidine kinase promoter, heat shock promoters, the early and late SV40 promoter, LTRs from retroviruses, and the mouse metallothionein-I promoter. Other promoters known to control expression of genes in prokaryotic or eukaryotic cells or their viruses may also be used.

Expression vectors and cloning vehicles

The invention provides expression vehicles comprising the chimeric compositions of the invention. These expression vehicles can be used in various screening assays. In one aspect, the nucleic acids of the invention comprise gene libraries, e.g., for screening for various intein constructs, enzyme domains or binding protein domains, detectable moiety domains, selection markers and the like.

Different vectors may be needed to suit the specific needs of different applications. In one aspect, the nucleic acids or expression vehicles (e.g., vectors) of the invention can comprise appropriate multiple cloning sites for fusing genes and/or libraries C-terminal to the intein-detectable domain fusion (e.g., intein-FP). In one aspect, the

5   nucleic acids or expression vehicles (e.g., vectors) of the invention can comprise promoters for inducible and/or constitutive expression of genes. The nucleic acids or expression vehicles (e.g., vectors) of the invention can be compatible with different screening hosts, such as bacilli, e.g., *E. coli, Pseudomonas, Streptomyces, Bacillus*, etc., yeast, insect, plant, mammalian and the like.

10   The invention provides expression vectors and cloning vehicles comprising nucleic acids of the invention, e.g., sequences encoding the inteins of the invention. Expression vectors and cloning vehicles of the invention can comprise viral particles, baculovirus, phage, plasmids, phagemids, cosmids, fosmids, bacterial artificial chromosomes, viral DNA (e.g., vaccinia, adenovirus, foul pox virus, pseudorabies and

15   derivatives of SV40), P1-based artificial chromosomes, yeast plasmids, yeast artificial chromosomes, and any other vectors specific for specific hosts of interest (such as bacillus, *Aspergillus* and yeast). Vectors of the invention can include chromosomal, non-chromosomal and synthetic DNA sequences. Large numbers of suitable vectors are known to those of skill in the art, and are commercially available. Exemplary vectors are

20   include: bacterial: pQE vectors (Qiagen), pBluescript plasmids, pNH vectors, (lambda-ZAP vectors (Stratagene); ptrc99a, pKK223-3, pDR540, pRIT2T (Pharmacia); Eukaryotic: pXT1, pSG5 (Stratagene), pSVK3, pBPV, pMSG, pSVLSV40 (Pharmacia). However, any other plasmid or other vector may be used so long as they are replicable and viable in the host. Low copy number or high copy number vectors may be employed

25   with the present invention.

The expression vector can comprise a promoter, a ribosome binding site for translation initiation and a transcription terminator. The vector may also include appropriate sequences for amplifying expression. Mammalian expression vectors can comprise an origin of replication, any necessary ribosome binding sites, a

30   polyadenylation site, splice donor and acceptor sites, transcriptional termination sequences, and 5' flanking non-transcribed sequences. In some aspects, DNA sequences derived from the SV40 splice and polyadenylation sites may be used to provide the required non-transcribed genetic elements.

In one aspect, the expression vectors contain one or more selectable marker genes to permit selection of host cells containing the vector. Such selectable markers include genes encoding dihydrofolate reductase or genes conferring neomycin resistance for eukaryotic cell culture, genes conferring tetracycline or ampicillin resistance in *E. coli*, and the *S. cerevisiae* TRP1 gene. Promoter regions can be selected from any desired gene using chloramphenicol transferase (CAT) vectors or other vectors with selectable markers.

Vectors for expressing the polypeptide or fragment thereof in eukaryotic cells can also contain enhancers to increase expression levels. Enhancers are cis-acting elements of DNA, usually from about 10 to about 300 bp in length that act on a promoter to increase its transcription. Examples include the SV40 enhancer on the late side of the replication origin bp 100 to 270, the cytomegalovirus early promoter enhancer, the polyoma enhancer on the late side of the replication origin, and the adenovirus enhancers.

Particular bacterial vectors which can be used include the commercially available plasmids comprising genetic elements of the well known cloning vector pBR322 (ATCC 37017), pKK223-3 (Pharmacia Fine Chemicals, Uppsala, Sweden), GEM1 (Promega Biotec, Madison, WI, USA) pQE70, pQE60, pQE-9 (Qiagen), pD10, psiX174 pBluescript II KS, pNH8A, pNH16a, pNH18A, pNH46A (Stratagene), ptrc99a, pKK223-3, pKK233-3, DR540, pRIT5 (Pharmacia), pKK232-8 and pCM7. Particular eukaryotic vectors include pSV2CAT, pOG44, pXT1, pSG (Stratagene) pSVK3, pBPV, pMSG, and pSVL (Pharmacia). However, any other vector may be used as long as it is replicable and viable in the host cell.

The nucleic acids of the invention can be expressed in expression cassettes, vectors or viruses and transiently or stably expressed in plant cells and seeds. One exemplary transient expression system uses episomal expression systems, e.g., cauliflower mosaic virus (CaMV) viral RNA generated in the nucleus by transcription of an episomal mini-chromosome containing supercoiled DNA, see, e.g., Covey (1990) Proc. Natl. Acad. Sci. USA 87:1633-1637. Alternatively, coding sequences, i.e., all or sub-fragments of sequences of the invention can be inserted into a plant host cell genome becoming an integral part of the host chromosomal DNA. Sense or antisense transcripts can be expressed in this manner. A vector comprising the sequences (e.g., promoters or coding regions) from nucleic acids of the invention can comprise a marker gene that confers a selectable phenotype on a plant cell or a seed. For example, the marker may encode biocide resistance, particularly antibiotic resistance, such as resistance to

kanamycin, G418, bleomycin, hygromycin, or herbicide resistance, such as resistance to chlorosulfuron or Basta.

Host cells and transformed cells

The invention also provides a transformed cell comprising a nucleic acid sequence of the invention, e.g., a sequence encoding an intein of the invention, or a vector of the invention. The host cell may be any of the host cells familiar to those skilled in the art, including prokaryotic cells, eukaryotic cells, such as bacterial cells, fungal cells, yeast cells, mammalian cells, insect cells, or plant cells. Exemplary bacterial cells include *E. coli*, *Streptomyces*, *Bacillus subtilis*, *Salmonella typhimurium* and various species within the genera *Pseudomonas*, *Streptomyces*, and *Staphylococcus*. Exemplary insect cells include *Drosophila S2* and *Spodoptera Sf9*. Exemplary animal cells include CHO, COS or Bowes melanoma or any mouse or human cell line. The selection of an appropriate host is within the abilities of those skilled in the art. Techniques for transforming a wide variety of higher plant species are well known and described in the technical and scientific literature. See, e.g., Weising (1988) Ann. Rev. Genet. 22:421-477, U.S. Patent No. 5,750,870.

The vector can be introduced into the host cells using any of a variety of techniques, including transformation, transfection, transduction, viral infection, gene guns, or Ti-mediated gene transfer. Particular methods include calcium phosphate transfection, DEAE-Dextran mediated transfection, lipofection, or electroporation (Davis, L., Dibner, M., Battey, I., Basic Methods in Molecular Biology, (1986)).

Cell-free translation systems can also be employed to produce a chimeric polypeptide of the invention. Cell-free translation systems can use mRNAs transcribed from a DNA construct comprising a promoter operably linked to a nucleic acid encoding the polypeptide or fragment thereof. In some aspects, the DNA construct may be linearized prior to conducting an in vitro transcription reaction. The transcribed mRNA is then incubated with an appropriate cell-free translation extract, such as a rabbit reticulocyte extract, to produce the desired polypeptide or fragment thereof.

The expression vectors can contain one or more selectable marker genes to provide a phenotypic trait for selection of transformed host cells such as dihydrofolate reductase or neomycin resistance for eukaryotic cell culture, or such as kanamycin, tetracycline or ampicillin resistance in *E. coli*.

Amplification of Nucleic Acids

In practicing the invention, nucleic acids of the invention and nucleic acids encoding the polypeptides of the invention, or modified nucleic acids of the invention, can be reproduced by amplification. Amplification can also be used to clone or modify the nucleic acids of the invention. Thus, the invention provides amplification primer sequence pairs for amplifying nucleic acids of the invention. In alternative aspects, where the primer pairs are capable of amplifying nucleic acid sequences including the exemplary SEQ ID NO:1, or a subsequence thereof

The exemplary SEQ ID NO:1 is

atggaaaaga cagaaaaaaa tgagcttgtc agaaaactca ttttcaaccc tcaaggagac   60

agggaggcga gcaaaaggaa gataataaag ggaaacccga caaacatatt tgaacttaac   120

gagataaagt attcctgggc ttttgacctt tacaagttaa tgggctttac aaacttctgg   180

ataccgaag agatacagat gcttgaagac aggaaacagt acgagaccgt tctatcagac   240

tacgaaaaga gggcatacga actcgtcctt tccttcctca tagcccttga ctccttttcaa   300

gtggacatgc ttaaagagtt cggaaggatg ataaccgccc ccgaagtaga aatggccata   360

acagctcagg aatttcagga atccgtccac gcgtactctt accagttcat actcgagtct   420

gtagttgatc cggttaaagc ggacgagagt tacaactact ggcgggagga tgaaagactt   480

ctggaaagga ataaagtaat agcagagctg tacaacgaat tcattagaaa acccaacgaa   540

gaaaacttta taaaggcaac aatagggaac tacatactcg agagcctgta cttttactct   600

ggatttgcct tcttctacac actgggaaga cagggcaaaa tgagaaacac tgtacagcaa   660

atcaaatata tcaacagggg tgagctctgc ttcattgagg gaacggaggt tttgacgaag   720

agggggttcg ttgatttcag ggagctgagg gaagacgatc ttgtagctca gtacgatata   780

gaaacagggg aaatttcctg gacaaaacct tacgcctacg ttgaaaggga ttacgagggt   840

tctatgtaca gattaaaaca tcctaaaagc aactgggaag tagtagctac tgaagggcac   900

gagttcatag taaggaacct gaaaacagga aaggagagaa aggaaccgat agaaaaggta   960

aaactacatc cctactctgc aattcccgtt gcgggaaggt acacgggaga agtggaagag   1020

tacgacctct gggaactcgt aagcggaaaa ggtataactc ttaaaacgag gagtgctgtg   1080

aagaataagt taacaccgat agaaaaactc ctgatagttc ttcaggcgga cgggacaata   1140

gacagtaaga gaaatggaaa gttcacaggc ttccaacaat taaagttctt cttcctcaaag   1200

tatagaaaga ttaacgagtt tgaaaaaaata ctcaatgaat gtgcaccta cggaattaaa   1260

tggaaaaagt acgagcgcca agacggaatt gcttacacag tttactatcc gaatgacctt   1320

ccgataaagc ctactaagtt ctttgacgaa tgggtgagac ttgatgagat aacggaagaa   1380

tggataaggg aatttgtgga agaactcgtc aagtgggacg gacacattcc gaaagacagg   1440

aataaaaaga aggtttatta ttactccaca aaagaaaaaa gaaacaagga ctttgtgcag   1500

21

gcactttgtg ctctgggagg tatgagaact gttgtcagta gagagagaaa tccgaaggcg   1560

aaaaaccccg tttacaggat atggatttac ctagaggacg actacataaa tacccaaaca   1620

atggtgaagg aagagttcta ctacaaaggt aaggtgtact gcgtgagcgt tcccaaaggg   1680

aacatagttg tgagatacaa agacagccgtt tgtattgccgg gcaactgcca cgttacgctc   1740

ttcaggaaca taataaacac actcaggaaa gaaaatcccg aattatttac gcctgagata   1800

gaaaagtgga tagtggagta cttcaagtac gcggtgaacg aagaaatcaa atgggggcag   1860

tatgttaccc agaaccagat actcggtatt aacgacgtct tgatagagag gtatataaag   1920

tatctcggaa acctgaggat tactcagatc ggctttgatc cgatatatcc agaggttaca   1980

gaaaacccct taaagtggat agacgagttt agaaagataa acaacactaa aacggacttc   2040

ttccaggcaa agcctcagac ctactcaaaa gccaacgaac tcaagtggta a   2091

Thus, an exemplary amplification primer sequence pair is residues 1 to 21 of SEQ ID NO:1 (atggaaaaga cagaaaaaaa t) and the complementary strand of the last 21 residues of SEQ ID NO:1 (gccaacgaac tcaagtggta a).

Amplification reactions can also be used to quantify the amount of nucleic acid in a sample (such as the amount of message in a cell sample), label the nucleic acid (e.g., to apply it to an array or a blot), detect the nucleic acid, or quantify the amount of a specific nucleic acid in a sample. In one aspect of the invention, message isolated from a cell or a cDNA library are amplified.

The skilled artisan can select and design suitable oligonucleotide amplification primers. Amplification methods are also well known in the art, and include, e.g., polymerase chain reaction, PCR (see, e.g., PCR PROTOCOLS, A GUIDE TO METHODS AND APPLICATIONS, ed. Innis, Academic Press, N.Y. (1990) and PCR STRATEGIES (1995), ed. Innis, Academic Press, Inc., N.Y., ligase chain reaction (LCR) (see, e.g., Wu (1989) Genomics 4:560; Landegren (1988) Science 241:1077; Barringer (1990) Gene 89:117; transcription amplification (see, e.g., Kwoh (1989) Proc. Natl. Acad. Sci. USA 86:1173); and, self-sustained sequence replication (see, e.g., Guatelli (1990) Proc. Natl. Acad. Sci. USA 87:1874); Q Beta replicase amplification (see, e.g., Smith (1997) J. Clin. Microbiol. 35:1477-1491), automated Q-beta replicase amplification assay (see, e.g., Burg (1996) Mol. Cell. Probes 10:257-271) and other RNA polymerase mediated techniques (e.g., NASBA, Cangene, Mississauga, Ontario); see also Berger (1987) Methods Enzymol. 152:307-316; Sambrook; Ausubel; U.S. Patent Nos. 4,683,195 and 4,683,202; Sooknanan (1995) Biotechnology 13:563-564.

22

Determining the degree of sequence identity

The invention provides nucleic acids having at least 90% sequence identity to SEQ ID NO:1. In one aspect, the invention provides nucleic acids and polypeptides having at least 99%, 98%, 97%, 96%, 95%, 90%, 85%, 80%, 75%, 70%, 65%, 60%, 55% or 50% sequence identity (homology) to SEQ ID NO:1. In alternative aspects, the sequence identity can be over a region of at least about 5, 10, 20, 30, 40, 50, 100, 150, 200, or more consecutive residues, or the full length of the nucleic acid or polypeptide.

The extent of sequence identity (homology) may be determined using any computer program and associated parameters, including those described herein, such as BLAST 2.2.2. or FASTA version 3.0t78, with the default parameters.

Homologous sequences also include RNA sequences in which uridines replace the thymines in the nucleic acid sequences. The homologous sequences may be obtained using any of the procedures described herein or may result from the correction of a sequencing error. It will be appreciated that the nucleic acid sequences as set forth herein can be represented in the traditional single character format (see, e.g., Stryer, Lubert. Biochemistry, 3rd Ed., W. H Freeman & Co., New York) or in any other format which records the identity of the nucleotides in a sequence.

Various sequence comparison programs identified herein are used in this aspect of the invention. Protein and/or nucleic acid sequence identities (homologies) may be evaluated using any of the variety of sequence comparison algorithms and programs known in the art. Such algorithms and programs include, but are not limited to, TBLASTN, BLASTP, FASTA, TFASTA, and CLUSTALW (Pearson and Lipman, Proc. Natl. Acad. Sci. USA 85(8):2444-2448, 1988; Altschul et al., J. Mol. Biol. 215(3):403-410, 1990; Thompson et al., Nucleic Acids Res. 22(2):4673-4680, 1994; Higgins et al., Methods Enzymol. 266:383-402, 1996; Altschul et al., J. Mol. Biol. 215(3):403-410, 1990; Altschul et al., Nature Genetics 3:266-272, 1993).

Homology or identity can be measured using sequence analysis software (e.g., Sequence Analysis Software Package of the Genetics Computer Group, University of Wisconsin Biotechnology Center, 1710 University Avenue, Madison, WI 53705). Such software matches similar sequences by assigning degrees of homology to various deletions, substitutions and other modifications. The terms "homology" and "identity" in the context of two or more nucleic acids or polypeptide sequences, refer to two or more sequences or subsequences that are the same or have a specified percentage of amino acid residues or nucleotides that are the same when compared and aligned for maximum

23

correspondence over a comparison window or designated region as measured using any number of sequence comparison algorithms or by manual alignment and visual inspection. For sequence comparison, one sequence can act as a reference sequence (a sequence of the invention, e.g., SEQ ID NO:1.

A "comparison window," as used herein, includes reference to a segment of any one of the numbers of contiguous residues. For example, in alternative aspects of the invention, contiguous residues ranging anywhere from 20 to the full length of an exemplary polypeptide or nucleic acid sequence of the invention are compared to a reference sequence of the same number of contiguous positions after the two sequences are optimally aligned. If the reference sequence has the requisite sequence identity to an exemplary polypeptide or nucleic acid sequence of the invention, e.g., 50%, 55%, 60%, 65%, 70%, 75%, 80%, 90% or 95% sequence identity to a sequence of the invention, including SEQ ID NO:1.

In alternative embodiments, various subsequences, e.g., ranging from about 20 to 100, about 50 to 200, and about 100 to 150, are compared to a reference sequence of the same number of contiguous positions after the two sequences are optimally aligned. Methods of alignment of sequence for comparison are well known in the art. Optimal alignment of sequences for comparison can be conducted, e.g., by the local homology algorithm of Smith & Waterman, Adv. Appl. Math. 2:482, 1981, by the homology alignment algorithm of Needleman & Wunsch, J. Mol. Biol. 48:443, 1970, by the search for similarity method of person & Lipman, Proc. Nat'l. Acad. Sci. USA 85:2444, 1988, by computerized implementations of these algorithms (GAP, BESTFIT, FASTA, and TFASTA in the Wisconsin Genetics Software Package, Genetics Computer Group, 575 Science Dr., Madison, WI), or by manual alignment and visual inspection. Other algorithms for determining homology or identity include, for example, in addition to a BLAST program (Basic Local Alignment Search Tool at the National Center for Biological Information), ALIGN, AMAS (Analysis of Multiply Aligned Sequences), AMPS (Protein Multiple Sequence Alignment), ASSET (Aligned Segment Statistical Evaluation Tool), BANDS, BESTSCOR, BIOSCAN (Biological Sequence Comparative Analysis Node), BLIMPS (BLocks IMProved Searcher), FASTA, Intervals & Points, BMB, CLUSTAL V, CLUSTAL W, CONSENSUS, LCONSENSUS, WCONSENSUS, Smith-Waterman algorithm, DARWIN, Las Vegas algorithm, FNAT (Forced Nucleotide Alignment Tool), Framealign, Framesearch, DYNAMIC, FILTER, FSAP (Fristensky Sequence Analysis Package), GAP (Global Alignment Program), GENAL, GIBBS,

24

GenQuest, ISSC (Sensitive Sequence Comparison), LALIGN (Local Sequence Alignment), LCP (Local Content Program), MACAW (Multiple Alignment Construction & Analysis Workbench), MAP (Multiple Alignment Program), MBLKP, MBLKN, PIMA (Pattern-Induced Multi-sequence Alignment), SAGA (Sequence Alignment by Genetic Algorithm) and WHAT-IF. Such alignment programs can also be used to screen genome databases to identify polynucleotide sequences having substantially identical sequences. A number of genome databases are available, for example, a substantial portion of the human genome is available as part of the Human Genome Sequencing Project (Gibbs, 1995). Several genomes have been sequenced, e.g., M. genitalium (Fraser et al., 1995), M. jannaschii (Bult et al., 1996), H. influenzae (Fleischmann et al., 1995), E. coli (Blattner et al., 1997), and yeast (S. cerevisiae) (Mewes et al., 1997), and D. melanogaster (Adams et al., 2000). Significant progress has also been made in sequencing the genomes of model organism, such as mouse, C. elegans, and Arabadopsis sp. Databases containing genomic information annotated with some functional information are maintained by different organization, and are accessible via the internet.

BLAST, BLAST 2.0 and BLAST 2.2.2 algorithms are also used to practice the invention. They are described, e.g., in Altschul (1977) Nuc. Acids Res. 25:3389-3402; Altschul (1990) J. Mol. Biol. 215:403-410. Software for performing BLAST analyses is publicly available through the National Center for Biotechnology Information. This algorithm involves first identifying high scoring sequence pairs (HSPs) by identifying short words of length W in the query sequence, which either match or satisfy some positive-valued threshold score T when aligned with a word of the same length in a database sequence. T is referred to as the neighborhood word score threshold (Altschul (1990) supra). These initial neighborhood word hits act as seeds for initiating searches to find longer HSPs containing them. The word hits are extended in both directions along each sequence for as far as the cumulative alignment score can be increased. Cumulative scores are calculated using, for nucleotide sequences, the parameters M (reward score for a pair of matching residues; always >0). For amino acid sequences, a scoring matrix is used to calculate the cumulative score. Extension of the word hits in each direction are halted when: the cumulative alignment score falls off by the quantity X from its maximum achieved value; the cumulative score goes to zero or below, due to the accumulation of one or more negative-scoring residue alignments; or the end of either sequence is reached. The BLAST algorithm parameters W, T, and X determine the sensitivity and speed of the alignment. The BLASTN program (for nucleotide sequences)

uses as defaults a wordlength (W) of 11, an expectation (E) of 10, M=5, N=-4 and a comparison of both strands. For amino acid sequences, the BLASTP program uses as defaults a wordlength of 3, and expectations (E) of 10, and the BLOSUM62 scoring matrix (see Henikoff & Henikoff (1989) Proc. Natl. Acad. Sci. USA 89:10915) alignments (B) of 50, expectation (E) of 10, M=5, N= -4, and a comparison of both strands. The BLAST algorithm also performs a statistical analysis of the similarity between two sequences (see, e.g., Karlin & Altschul (1993) Proc. Natl. Acad. Sci. USA 90:5873). One measure of similarity provided by BLAST algorithm is the smallest sum probability (P(N)), which provides an indication of the probability by which a match between two nucleotide or amino acid sequences would occur by chance. For example, a nucleic acid is considered similar to a references sequence if the smallest sum probability in a comparison of the test nucleic acid to the reference nucleic acid is less than about 0.2, more preferably less than about 0.01, and most preferably less than about 0.001. In one aspect, protein and nucleic acid sequence homologies are evaluated using the Basic Local Alignment Search Tool ("BLAST"). For example, five specific BLAST programs can be used to perform the following task: (1) BLASTP and BLAST3 compare an amino acid query sequence against a protein sequence database; (2) BLASTN compares a nucleotide query sequence against a nucleotide sequence database; (3) BLASTX compares the six-frame conceptual translation products of a query nucleotide sequence (both strands) against a protein sequence database; (4) TBLASTN compares a query protein sequence against a nucleotide sequence database translated in all six reading frames (both strands); and, (5) TBLASTX compares the six-frame translations of a nucleotide query sequence against the six-frame translations of a nucleotide sequence database. The BLAST programs identify homologous sequences by identifying similar segments, which are referred to herein as "high-scoring segment pairs," between a query amino or nucleic acid sequence and a test sequence which is preferably obtained from a protein or nucleic acid sequence database. High-scoring segment pairs are preferably identified (i.e., aligned) by means of a scoring matrix, many of which are known in the art. Preferably, the scoring matrix used is the BLOSUM62 matrix (Gonnet et al., Science 256:1443-1445, 1992; Henikoff and Henikoff, Proteins 17:49-61, 1993). Less preferably, the PAM or PAM250 matrices may also be used (see, e.g., Schwartz and Dayhoff, eds., 1978, Matrices for Detecting Distance Relationships: Atlas of Protein Sequence and Structure, Washington: National Biomedical Research Foundation).

In one aspect of the invention, to determine if a nucleic acid has the requisite sequence identity to be within the scope of the invention, the NCBI BLAST 2.2.2 programs is used, default options to blastp. There are about 38 setting options in the BLAST 2.2.2 program. In this exemplary aspect of the invention, all default values are used except for the default filtering setting (i.e., all parameters set to default except filtering which is set to OFF); in its place a "-F F" setting is used, which disables filtering. Use of default filtering often results in Karlin-Altschul violations due to short length of sequence.

The default values used in this exemplary aspect of the invention include:

"Filter for low complexity: ON

Word Size: 3

Matrix: Blosum62

Gap Costs: Existence:11

Extension:1"

Other default settings can be: filter for low complexity OFF, word size of 3 for protein, BLOSUM62 matrix, gap existence penalty of -11 and a gap extension penalty of -1. An exemplary NCBI BLAST 2.2.2 program setting has the "-W" option default to 0. This means that, if not set, the word size defaults to 3 for proteins and 11 for nucleotides.

Computer systems and computer program products

To determine and identify sequence identities, structural homologies, motifs and the like *in silico*, the sequence of the invention can be stored, recorded, and manipulated on any medium which can be read and accessed by a computer. Accordingly, the invention provides computers, computer systems, computer readable mediums, computer programs products and the like recorded or stored thereon the nucleic acid and polypeptide sequences of the invention. As used herein, the words "recorded" and "stored" refer to a process for storing information on a computer medium. A skilled artisan can readily adopt any known methods for recording information on a computer readable medium to generate manufactures comprising one or more of the nucleic acid and/or polypeptide sequences of the invention.

Another aspect of the invention is a computer readable medium having recorded thereon at least one nucleic acid and/or polypeptide sequence of the invention. Computer readable media include magnetically readable media, optically readable media,

electronically readable media and magnetic/optical media. For example, the computer
readable media may be a hard disk, a floppy disk, a magnetic tape, CD-ROM, Digital
Versatile Disk (DVD), Random Access Memory (RAM), or Read Only Memory (ROM)
as well as other types of other media known to those skilled in the art.

5   Hybridization of nucleic acids
                    The invention provides isolated or recombinant nucleic acids that
hybridize under stringent conditions to an exemplary sequence of the invention, e.g., a
sequence as set forth in SEQ ID NO:1, or a nucleic acid that encodes a polypeptide of the
invention, e.g., SEQ ID NO:2. The stringent conditions can be highly stringent
10   conditions, medium stringent conditions, low stringent conditions, including the high and
reduced stringency conditions described herein. In one aspect, it is the stringency of the
wash conditions that set forth the conditions which determine whether a nucleic acid is
within the scope of the invention, as discussed below.
                    In alternative embodiments, nucleic acids of the invention as defined by
15   their ability to hybridize under stringent conditions can be between about five residues
and the full length of nucleic acid of the invention; e.g., they can be at least 5, 10, 15, 20,
25, 30, 35, 40, 50, 55, 60, 65, 70, 75, 80, 90, 100, 150, 200, 250, or more, residues in
length. Nucleic acids shorter than full length are also included. These nucleic acids can
be useful as, e.g., hybridization probes, labeling probes, PCR oligonucleotide probes,
20   iRNA, antisense or sequences encoding antibody binding peptides (epitopes), motifs,
active sites and the like.
                    In one aspect, nucleic acids of the invention are defined by their ability to
hybridize under high stringency comprises conditions of about 50% formamide at about
37°C to 42°C. In one aspect, nucleic acids of the invention are defined by their ability to
25   hybridize under reduced stringency comprising conditions in about 35% to 25%
formamide at about 30°C to 35°C.
                    Alternatively, nucleic acids of the invention are defined by their ability to
hybridize under high stringency comprising conditions at 42°C in 50% formamide, 5X
SSPE, 0.3% SDS, and a repetitive sequence blocking nucleic acid, such as cot-1 or
30   salmon sperm DNA (e.g., 200 n/ml sheared and denatured salmon sperm DNA). In one
aspect, nucleic acids of the invention are defined by their ability to hybridize under
reduced stringency conditions comprising 35% formamide at a reduced temperature of
35°C. Following hybridization, the filter may be washed with 6X SSC, 0.5% SDS at

50°C. These conditions are considered to be "moderate" conditions above 25% formamide and "low" conditions below 25% formamide. A specific example of "moderate" hybridization conditions is when the above hybridization is conducted at 30% formamide. A specific example of "low stringency" hybridization conditions is when the above hybridization is conducted at 10% formamide.

The temperature range corresponding to a particular level of stringency can be further narrowed by calculating the purine to pyrimidine ratio of the nucleic acid of interest and adjusting the temperature accordingly. Nucleic acids of the invention are also defined by their ability to hybridize under high, medium, and low stringency conditions as set forth in Ausubel and Sambrook. Variations on the above ranges and conditions are well known in the art. Hybridization conditions are discussed further, below.

The above procedure may be modified to identify nucleic acids having decreasing levels of homology to the probe sequence. For example, to obtain nucleic acids of decreasing homology to the detectable probe, less stringent conditions may be used. For example, the hybridization temperature may be decreased in increments of 5°C from 68°C to 42°C in a hybridization buffer having a Na^+ concentration of approximately 1M. Following hybridization, the filter may be washed with 2X SSC, 0.5% SDS at the temperature of hybridization. These conditions are considered to be "moderate" conditions above 50°C and "low" conditions below 50°C. A specific example of "moderate" hybridization conditions is when the above hybridization is conducted at 55°C. A specific example of "low stringency" hybridization conditions is when the above hybridization is conducted at 45°C.

Alternatively, the hybridization may be carried out in buffers, such as 6X SSC, containing formamide at a temperature of 42°C. In this case, the concentration of formamide in the hybridization buffer may be reduced in 5% increments from 50% to 0% to identify clones having decreasing levels of homology to the probe. Following hybridization, the filter may be washed with 6X SSC, 0.5% SDS at 50°C. These conditions are considered to be "moderate" conditions above 25% formamide and "low" conditions below 25% formamide. A specific example of "moderate" hybridization conditions is when the above hybridization is conducted at 30% formamide. A specific example of "low stringency" hybridization conditions is when the above hybridization is conducted at 10% formamide.

29

However, the selection of a hybridization format is not critical - it is the stringency of the wash conditions that set forth the conditions which determine whether a nucleic acid is within the scope of the invention. Wash conditions used to identify nucleic acids within the scope of the invention include, e.g.: a salt concentration of about 0.02 molar at pH 7 and a temperature of at least about 50°C or about 55°C to about 60°C; or, a salt concentration of about 0.15 M NaCl at 72°C for about 15 minutes; or, a salt concentration of about 0.2X SSC at a temperature of at least about 50°C or about 55°C to about 60°C for about 15 to about 20 minutes; or, the hybridization complex is washed twice with a solution with a salt concentration of about 2X SSC containing 0.1% SDS at room temperature for 15 minutes and then washed twice by 0.1X SSC containing 0.1% SDS at 68oC for 15 minutes; or, equivalent conditions. See Sambrook, Tijssen and Ausubel for a description of SSC buffer and equivalent conditions.

These methods may be used to isolate nucleic acids of the invention.

Oligonucleotides probes and methods for using them

The invention also provides nucleic acid probes that can be used, e.g., for identifying nucleic acids encoding an intein of the invention. In one aspect, the probe comprises at least 10 consecutive bases of a nucleic acid of the invention. Alternatively, a probe of the invention can be at least about 5, 6, 7, 8, 9, 10, 15, 20, 25, 30, 35, 40, 45, 50, 60, 70, 80, 90, 100, 110, 120, 130, 150 or about 10 to 50, about 20 to 60 about 30 to 70, consecutive bases of a sequence as set forth in a nucleic acid of the invention. The probes identify a nucleic acid by binding and/or hybridization. The probes can be used in arrays of the invention, see discussion below, including, e.g., capillary arrays. The probes of the invention can also be used to isolate other nucleic acids or polypeptides.

The probes of the invention can be used to determine whether a biological sample, such as a soil sample, contains an organism having a nucleic acid sequence of the invention or an organism from which the nucleic acid was obtained. In such procedures, a biological sample potentially harboring the organism from which the nucleic acid was isolated is obtained and nucleic acids are obtained from the sample. The nucleic acids are contacted with the probe under conditions which permit the probe to specifically hybridize to any complementary sequences present in the sample. Where necessary, conditions which permit the probe to specifically hybridize to complementary sequences may be determined by placing the probe in contact with complementary sequences from samples known to contain the complementary sequence, as well as control sequences

which do not contain the complementary sequence. Hybridization conditions, such as the salt concentration of the hybridization buffer, the formamide concentration of the hybridization buffer, or the hybridization temperature, may be varied to identify conditions which allow the probe to hybridize specifically to complementary nucleic acids (see discussion on specific hybridization conditions).

Modification of Nucleic Acids

The invention provides methods for mutating a nucleic acid sequence of the invention, including an intein of the invention. In one aspect the methods comprise providing segments of a chromosome or part of a chromosome and introducing one or more mutations into one or more of the segments. Any method can be used to introduce the mutation, either randomly, non-randomly, or both. These methods can be repeated or used in various combinations to generate one or more mutations into one or more of the segments. These methods also can be repeated or used in various combinations. In another aspect, the genetic composition of a cell is altered by, e.g., modification of a homologous gene *ex vivo* by a method of the invention followed by its reinsertion into a cell.

Any method can be used to introduce the mutation, either randomly, non-randomly, or both. For example, random or stochastic methods, or, non-stochastic, or "directed evolution," methods, see, e.g., U.S. Patent No. 6,361,974. Methods for random mutation of genes are well known in the art, see, e.g., U.S. Patent No. 5,830,696. For example, mutagens can be used to randomly mutate a gene. Mutagens include, e.g., ultraviolet light or gamma irradiation, or a chemical mutagen, e.g., mitomycin, nitrous acid, photoactivated psoralens, alone or in combination, to induce DNA breaks amenable to repair by recombination. Other chemical mutagens include, for example, sodium bisulfite, nitrous acid, hydroxylamine, hydrazine or formic acid. Other mutagens are analogues of nucleotide precursors, e.g., nitrosoguanidine, 5-bromouracil, 2-aminopurine, or acridine. These agents can be added to a PCR reaction in place of the nucleotide precursor thereby mutating the sequence. Intercalating agents such as proflavine, acriflavine, quinacrine and the like can also be used.

Any technique in molecular biology can be used, e.g., random PCR mutagenesis, see, e.g., Rice (1992) Proc. Natl. Acad. Sci. USA 89:5467-5471; or, combinatorial multiple cassette mutagenesis, see, e.g., Crameri (1995) Biotechniques 18:194-196. Alternatively, nucleic acids, e.g., genes, can be reassembled after random, or

"stochastic," fragmentation, see, e.g., U.S. Patent Nos. 6,291,242; 6,287,862; 6,287,861;
5,955,358; 5,830,721; 5,824,514; 5,811,238; 5,605,793. In alternative aspects,
modifications, additions or deletions are introduced by error-prone PCR, shuffling,
oligonucleotide-directed mutagenesis, assembly PCR, sexual PCR mutagenesis, in vivo
mutagenesis, cassette mutagenesis, recursive ensemble mutagenesis, exponential
ensemble mutagenesis, site-specific mutagenesis, gene reassembly, gene site saturated
mutagenesis (GSSM), synthetic ligation reassembly (SLR), recombination, recursive
sequence recombination, phosphothioate-modified DNA mutagenesis, uracil-containing
template mutagenesis, gapped duplex mutagenesis, point mismatch repair mutagenesis,
repair-deficient host strain mutagenesis, chemical mutagenesis, radiogenic mutagenesis,
deletion mutagenesis, restriction-selection mutagenesis, restriction-purification
mutagenesis, artificial gene synthesis, ensemble mutagenesis, chimeric nucleic acid
multimer creation, and/or a combination of these and other methods.

       The following publications describe a variety of recursive recombination
procedures and/or methods which can be incorporated into the methods of the invention:
Stemmer (1999) "Molecular breeding of viruses for targeting and other clinical
properties" Tumor Targeting 4:1-4; Ness (1999) Nature Biotechnology 17:893-896;
Chang (1999) "Evolution of a cytokine using DNA family shuffling" Nature
Biotechnology 17:793-797; Minshull (1999) "Protein evolution by molecular breeding"
Current Opinion in Chemical Biology 3:284-290; Christians (1999) "Directed evolution
of thymidine kinase for AZT phosphorylation using DNA family shuffling" Nature
Biotechnology 17:259-264; Crameri (1998) "DNA shuffling of a family of genes from
diverse species accelerates directed evolution" Nature 391:288-291; Crameri (1997)
"Molecular evolution of an arsenate detoxification pathway by DNA shuffling," Nature
Biotechnology 15:436-438; Zhang (1997) "Directed evolution of an effective fucosidase
from a galactosidase by DNA shuffling and screening" Proc. Natl. Acad. Sci. USA
94:4504-4509; Patten et al. (1997) "Applications of DNA Shuffling to Pharmaceuticals
and Vaccines" Current Opinion in Biotechnology 8:724-733; Crameri et al. (1996)
"Construction and evolution of antibody-phage libraries by DNA shuffling" Nature
Medicine 2:100-103; Gates et al. (1996) "Affinity selective isolation of ligands from
peptide libraries through display on a lac repressor 'headpiece dimer'" Journal of
Molecular Biology 255:373-386; Stemmer (1996) "Sexual PCR and Assembly PCR" In:
The Encyclopedia of Molecular Biology. VCH Publishers, New York. pp.447-457;
Crameri and Stemmer (1995) "Combinatorial multiple cassette mutagenesis creates all the

permutations of mutant and wildtype cassettes" BioTechniques 18:194-195; Stemmer et

al. (1995) "Single-step assembly of a gene and entire plasmid form large numbers of

oligodeoxyribonucleotides" Gene, 164:49-53; Stemmer (1995) "The Evolution of

Molecular Computation" Science 270: 1510; Stemmer (1995) "Searching Sequence

Space" Bio/Technology 13:549-553; Stemmer (1994) "Rapid evolution of a protein in

vitro by DNA shuffling" Nature 370:389-391; and Stemmer (1994) "DNA shuffling by

random fragmentation and reassembly: In vitro recombination for molecular evolution."

Proc. Natl. Acad. Sci. USA 91:10747-10751.

       Mutational methods of generating diversity include, for example, site-

directed mutagenesis (Ling et al. (1997) "Approaches to DNA mutagenesis: an overview"

Anal Biochem. 254(2): 157-178; Dale et al. (1996) "Oligonucleotide-directed random

mutagenesis using the phosphorothioate method" Methods Mol. Biol. 57:369-374; Smith

(1985) "In vitro mutagenesis" Ann. Rev. Genet. 19:423-462; Botstein & Shortle (1985)

"Strategies and applications of in vitro mutagenesis" Science 229:1193-1201; Carter

(1986) "Site-directed mutagenesis" Biochem. J. 237:1-7; and Kunkel (1987) "The

efficiency of oligonucleotide directed mutagenesis" in Nucleic Acids & Molecular

Biology (Eckstein, F. and Lilley, D. M. J. eds., Springer Verlag, Berlin)); mutagenesis

using uracil containing templates (Kunkel (1985) "Rapid and efficient site-specific

mutagenesis without phenotypic selection" Proc. Natl. Acad. Sci. USA 82:488-492;

Kunkel et al. (1987) "Rapid and efficient site-specific mutagenesis without phenotypic

selection" Methods in Enzymol. 154, 367-382; and Bass et al. (1988) "Mutant Trp

repressors with new DNA-binding specificities" Science 242:240-245); oligonucleotide-

directed mutagenesis (Methods in Enzymol. 100: 468-500 (1983); Methods in Enzymol.

154: 329-350 (1987); Zoller & Smith (1982) "Oligonucleotide-directed mutagenesis using

M13-derived vectors: an efficient and general procedure for the production of point

mutations in any DNA fragment" Nucleic Acids Res. 10:6487-6500; Zoller & Smith

(1983) "Oligonucleotide-directed mutagenesis of DNA fragments cloned into M13

vectors" Methods in Enzymol. 100:468-500; and Zoller & Smith (1987) Oligonucleotide-

directed mutagenesis: a simple method using two oligonucleotide primers and a single-

stranded DNA template" Methods in Enzymol. 154:329-350); phosphorothioate-modified

DNA mutagenesis (Taylor et al. (1985) "The use of phosphorothioate-modified DNA in

restriction enzyme reactions to prepare nicked DNA" Nucl. Acids Res. 13: 8749-8764;

Taylor et al. (1985) "The rapid generation of oligonucleotide-directed mutations at high

frequency using phosphorothioate-modified DNA" Nucl. Acids Res. 13: 8765-8787

(1985); Nakamaye (1986) "Inhibition of restriction endonuclease Nci I cleavage by phosphorothioate groups and its application to oligonucleotide-directed mutagenesis" Nucl. Acids Res. 14: 9679-9598; Sayers et al. (1988) "Y-T Exonucleases in phosphorothioate-based oligonucleotide-directed mutagenesis" Nucl. Acids Res. 16:791-802; and Sayers et al. (1988) "Strand specific cleavage of phosphorothioate-containing DNA by reaction with restriction endonucleases in the presence of ethidium bromide" Nucl. Acids Res. 16: 803-814); mutagenesis using gapped duplex DNA (Kramer et al. (1984) "The gapped duplex DNA approach to oligonucleotide-directed mutation construction" Nucl. Acids Res. 12: 9441-9456; Kramer & Fritz (1987) Methods in Enzymol. "Oligonucleotide-directed construction of mutations via gapped duplex DNA" 154:350-367; Kramer et al. (1988) "Improved enzymatic in vitro reactions in the gapped duplex DNA approach to oligonucleotide-directed construction of mutations" Nucl. Acids Res. 16: 7207; and Fritz et al. (1988) "Oligonucleotide-directed construction of mutations: a gapped duplex DNA procedure without enzymatic reactions in vitro" Nucl. Acids Res. 16: 6987-6999).

Additional protocols used in the methods of the invention include point mismatch repair (Kramer (1984) "Point Mismatch Repair" Cell 38:879-887), mutagenesis using repair-deficient host strains (Carter et al. (1985) "Improved oligonucleotide site-directed mutagenesis using M13 vectors" Nucl. Acids Res. 13: 4431-4443; and Carter (1987) "Improved oligonucleotide-directed mutagenesis using M13 vectors" Methods in Enzymol. 154: 382-403), deletion mutagenesis (Eghtedarzadeh (1986) "Use of oligonucleotides to generate large deletions" Nucl. Acids Res. 14: 5115), restriction-selection and restriction-selection and restriction-purification (Wells et al. (1986) "Importance of hydrogen-bond formation in stabilizing the transition state of subtilisin" Phil. Trans. R. Soc. Lond. A 317: 415-423), mutagenesis by total gene synthesis (Nambiar et al. (1984) "Total synthesis and cloning of a gene coding for the ribonuclease S protein" Science 223: 1299-1301; Sakamar and Khorana (1988) "Total synthesis and expression of a gene for the a-subunit of bovine rod outer segment guanine nucleotide-binding protein (transducin)" Nucl. Acids Res. 14: 6361-6372; Wells et al. (1985) "Cassette mutagenesis: an efficient method for generation of multiple mutations at defined sites" Gene 34:315-323; and Grundstrom et al. (1985) "Oligonucleotide-directed mutagenesis by microscale `shot-gun` gene synthesis" Nucl. Acids Res. 13: 3305-3316), double-strand break repair (Mandecki (1986); Arnold (1993) "Protein engineering for unusual environments" Current Opinion in Biotechnology 4:450-455. "Oligonucleotide-

directed double-strand break repair in plasmids of Escherichia coli: a method for site-specific mutagenesis" Proc. Natl. Acad. Sci. USA, 83:7177-7181). Additional details on many of the above methods can be found in Methods in Enzymology Volume 154, which also describes useful controls for trouble-shooting problems with various mutagenesis methods.

Additional protocols used in the methods of the invention include those discussed in U.S. Patent Nos. 5,605,793 to Stemmer (Feb. 25, 1997), "Methods for In Vitro Recombination;" U.S. Pat. No. 5,811,238 to Stemmer et al. (Sep. 22, 1998) "Methods for Generating Polynucleotides having Desired Characteristics by Iterative Selection and Recombination;" U.S. Pat. No. 5,830,721 to Stemmer et al. (Nov. 3, 1998), "DNA Mutagenesis by Random Fragmentation and Reassembly;" U.S. Pat. No. 5,834,252 to Stemmer, et al. (Nov. 10, 1998) "End-Complementary Polymerase Reaction;" U.S. Pat. No. 5,837,458 to Minshull, et al. (Nov. 17, 1998), "Methods and Compositions for Cellular and Metabolic Engineering;" WO 95/22625, Stemmer and Crameri, "Mutagenesis by Random Fragmentation and Reassembly;" WO 96/33207 by Stemmer and Lipschutz "End Complementary Polymerase Chain Reaction;" WO 97/20078 by Stemmer and Crameri "Methods for Generating Polynucleotides having Desired Characteristics by Iterative Selection and Recombination;" WO 97/35966 by Minshull and Stemmer, "Methods and Compositions for Cellular and Metabolic Engineering;" WO 99/41402 by Punnonen et al. "Targeting of Genetic Vaccine Vectors;" WO 99/41383 by Punnonen et al. "Antigen Library Immunization;" WO 99/41369 by Punnonen et al. "Genetic Vaccine Vector Engineering;" WO 99/41368 by Punnonen et al. "Optimization of Immunomodulatory Properties of Genetic Vaccines;" EP 752008 by Stemmer and Crameri, "DNA Mutagenesis by Random Fragmentation and Reassembly;" EP 0932670 by Stemmer "Evolving Cellular DNA Uptake by Recursive Sequence Recombination;" WO 99/23107 by Stemmer et al., "Modification of Virus Tropism and Host Range by Viral Genome Shuffling;" WO 99/21979 by Apt et al., "Human Papillomavirus Vectors;" WO 98/31837 by del Cardayre et al. "Evolution of Whole Cells and Organisms by Recursive Sequence Recombination;" WO 98/27230 by Patten and Stemmer, "Methods and Compositions for Polypeptide Engineering;" WO 98/27230 by Stemmer et al., "Methods for Optimization of Gene Therapy by Recursive Sequence Shuffling and Selection," WO 00/00632, "Methods for Generating Highly Diverse Libraries," WO 00/09679, "Methods for Obtaining in Vitro Recombined Polynucleotide Sequence Banks and Resulting Sequences," WO 98/42832 by Arnold et al., "Recombination of

Polynucleotide Sequences Using Random or Defined Primers," WO 99/29902 by Arnold et al., "Method for Creating Polynucleotide and Polypeptide Sequences," WO 98/41653 by Vind, "An in Vitro Method for Construction of a DNA Library," WO 98/41622 by Borchert et al., "Method for Constructing a Library Using DNA Shuffling," and WO 98/42727 by Pati and Zarling, "Sequence Alterations using Homologous Recombination."

Protocols that can be used to practice the invention (providing details regarding various diversity generating methods) are described, e.g., in U.S. Patent application serial no. (USSN) 09/407,800, "SHUFFLING OF CODON ALTERED GENES" by Patten et al. filed Sep. 28, 1999; "EVOLUTION OF WHOLE CELLS AND ORGANISMS BY RECURSIVE SEQUENCE RECOMBINATION" by del Cardayre et al., United States Patent No. 6,379,964; "OLIGONUCLEOTIDE MEDIATED NUCLEIC ACID RECOMBINATION" by Crameri et al., United States Patent Nos. 6,319,714; 6,368,861; 6,376,246; 6,423,542; 6,426,224 and PCT/US00/01203; "USE OF CODON-VARIED OLIGONUCLEOTIDE SYNTHESIS FOR SYNTHETIC SHUFFLING" by Welch et al., United States Patent No. 6,436,675; "METHODS FOR MAKING CHARACTER STRINGS, POLYNUCLEOTIDES & POLYPEPTIDES HAVING DESIRED CHARACTERISTICS" by Selifonov et al., filed Jan. 18, 2000, (PCT/US00/01202) and, e.g. "METHODS FOR MAKING CHARACTER STRINGS, POLYNUCLEOTIDES & POLYPEPTIDES HAVING DESIRED CHARACTERISTICS" by Selifonov et al., filed Jul. 18, 2000 (U.S. Ser. No. 09/618,579); "METHODS OF POPULATING DATA STRUCTURES FOR USE IN EVOLUTIONARY SIMULATIONS" by Selifonov and Stemmer, filed Jan. 18, 2000 (PCT/US00/01138); and "SINGLE-STRANDED NUCLEIC ACID TEMPLATE-MEDIATED RECOMBINATION AND NUCLEIC ACID FRAGMENT ISOLATION" by Affholter, filed Sep. 6, 2000 (U.S. Ser. No. 09/656,549); and United States Patent Nos. 6,177,263; 6,153,410.

Non-stochastic, or "directed evolution," methods include, e.g., saturation mutagenesis (GSSM), synthetic ligation reassembly (SLR), or a combination thereof are used to modify the nucleic acids of the invention to generate mutated nucleic acid segments. Polypeptides encoded by these modified nucleic acids can be screened for an activity before testing for proteolytic or other activity. Any testing modality or protocol can be used, e.g., using a capillary array platform. See, e.g., U.S. Patent Nos. 6,361,974; 6,280,926; 5,939,250.

Saturation mutagenesis, or, GSSM

In one aspect of the invention, non-stochastic gene modification, a "directed evolution process," is used to generate mutated sequences. Variations of this method have been termed "gene site-saturation mutagenesis," "site-saturation mutagenesis," "saturation mutagenesis" or simply "GSSM." It can be used in combination with other mutagenization processes. See, e.g., U.S. Patent Nos. 6,171,820; 6,238,884. In one aspect, GSSM comprises providing a template polynucleotide and a plurality of oligonucleotides, wherein each oligonucleotide comprises a sequence homologous to the template polynucleotide, thereby targeting a specific sequence of the template polynucleotide, and a sequence that is a variant of the homologous gene; generating progeny polynucleotides comprising non-stochastic sequence variations by replicating the template polynucleotide with the oligonucleotides, thereby generating polynucleotides comprising homologous gene sequence variations.

In one aspect, codon primers containing a degenerate N,N,G/T sequence are used to introduce point mutations into a polynucleotide, so as to generate a set of progeny polypeptides in which a full range of single amino acid substitutions is represented at each amino acid position, e.g., an amino acid residue in an enzyme active site or ligand binding site targeted to be modified. These oligonucleotides can comprise a contiguous first homologous sequence, a degenerate N,N,G/T sequence, and, optionally, a second homologous sequence. The downstream progeny translational products from the use of such oligonucleotides include all possible amino acid changes at each amino acid site along the polypeptide, because the degeneracy of the N,N,G/T sequence includes codons for all 20 amino acids. In one aspect, one such degenerate oligonucleotide (comprised of, e.g., one degenerate N,N,G/T cassette) is used for subjecting each original codon in a parental polynucleotide template to a full range of codon substitutions. In another aspect, at least two degenerate cassettes are used – either in the same oligonucleotide or not, for subjecting at least two original codons in a parental polynucleotide template to a full range of codon substitutions. For example, more than one N,N,G/T sequence can be contained in one oligonucleotide to introduce amino acid mutations at more than one site. This plurality of N,N,G/T sequences can be directly contiguous, or separated by one or more additional nucleotide sequence(s). In another aspect, oligonucleotides serviceable for introducing additions and deletions can be used either alone or in combination with the codons containing an N,N,G/T sequence, to

37

introduce any combination or permutation of amino acid additions, deletions, and/or
substitutions.

In one aspect, simultaneous mutagenesis of two or more contiguous amino
acid positions is done using an oligonucleotide that contains contiguous N,N,G/T triplets,
i.e. a degenerate (N,N,G/T)n sequence. In another aspect, degenerate cassettes having
less degeneracy than the N,N,G/T sequence are used. For example, it may be desirable in
some instances to use (e.g. in an oligonucleotide) a degenerate triplet sequence comprised
of only one N, where said N can be in the first second or third position of the triplet. Any
other bases including any combinations and permutations thereof can be used in the
remaining two positions of the triplet. Alternatively, it may be desirable in some
instances to use (e.g. in an oligo) a degenerate N,N,N triplet sequence.

In one aspect, use of degenerate triplets (e.g., N,N,G/T triplets) allows for
systematic and easy generation of a full range of possible natural amino acids (for a total
of 20 amino acids) into each and every amino acid position in a polypeptide (in
alternative aspects, the methods also include generation of less than all possible
substitutions per amino acid residue, or codon, position). For example, for a 100 amino
acid polypeptide, 2000 distinct species (i.e. 20 possible amino acids per position X 100
amino acid positions) can be generated. Through the use of an oligonucleotide or set of
oligonucleotides containing a degenerate N,N,G/T triplet, 32 individual sequences can
code for all 20 possible natural amino acids. Thus, in a reaction vessel in which a
parental polynucleotide sequence is subjected to saturation mutagenesis using at least one
such oligonucleotide, there are generated 32 distinct progeny polynucleotides encoding
20 distinct polypeptides. In contrast, the use of a non-degenerate oligonucleotide in site-
directed mutagenesis leads to only one progeny polypeptide product per reaction vessel.
Nondegenerate oligonucleotides can optionally be used in combination with degenerate
primers disclosed; for example, nondegenerate oligonucleotides can be used to generate
specific point mutations in a working polynucleotide. This provides one means to
generate specific silent point mutations, point mutations leading to corresponding amino
acid changes, and point mutations that cause the generation of stop codons and the
corresponding expression of polypeptide fragments.

In one aspect, each saturation mutagenesis reaction vessel contains
polynucleotides encoding at least 20 progeny polypeptide molecules such that all 20
natural amino acids are represented at the one specific amino acid position corresponding
to the codon position mutagenized in the parental polynucleotide (other aspects use less

than all 20 natural combinations). The 32-fold degenerate progeny polypeptides generated from each saturation mutagenesis reaction vessel can be subjected to clonal amplification (e.g. cloned into a suitable host, e.g., E. coli host, using, e.g., an expression vector) and subjected to expression screening. When an individual progeny polypeptide is identified by screening to display a favorable change in property (when compared to the parental polypeptide, such as increased proteolytic activity under alkaline or acidic conditions), it can be sequenced to identify the correspondingly favorable amino acid substitution contained therein.

In one aspect, upon mutagenizing each and every amino acid position in a parental polypeptide using saturation mutagenesis as disclosed herein, favorable amino acid changes may be identified at more than one amino acid position. One or more new progeny molecules can be generated that contain a combination of all or part of these favorable amino acid substitutions. For example, if 2 specific favorable amino acid changes are identified in each of 3 amino acid positions in a polypeptide, the permutations include 3 possibilities at each position (no change from the original amino acid, and each of two favorable changes) and 3 positions. Thus, there are 3 x 3 x 3 or 27 total possibilities, including 7 that were previously examined - 6 single point mutations (i.e. 2 at each of three positions) and no change at any position.

In another aspect, site-saturation mutagenesis can be used together with another stochastic or non-stochastic means to vary sequence, e.g., synthetic ligation reassembly (see below), shuffling, chimerization, recombination and other mutagenizing processes and mutagenizing agents. This invention provides for the use of any mutagenizing process(es), including saturation mutagenesis, in an iterative manner.

Synthetic Ligation Reassembly (SLR)

The methods of the invention include use of non-stochastic gene modification system termed "synthetic ligation reassembly," or simply "SLR," a "directed evolution process," to generate mutated sequences of the invention. SLR is a method of ligating oligonucleotide fragments together non-stochastically. This method differs from stochastic oligonucleotide shuffling in that the nucleic acid building blocks are not shuffled, concatenated or chimerized randomly, but rather are assembled non-stochastically. See, e.g., U.S. Patent Application Serial No. (USSN) 09/332,835 entitled "Synthetic Ligation Reassembly in Directed Evolution" and filed on June 14, 1999 ("USSN 09/332,835"). In one aspect, SLR comprises the following steps: (a) providing a

template polynucleotide, wherein the template polynucleotide comprises sequence encoding a homologous gene; (b) providing a plurality of building block polynucleotides, wherein the building block polynucleotides are designed to cross-over reassemble with the template polynucleotide at a predetermined sequence, and a building block

5   polynucleotide comprises a sequence that is a variant of the homologous gene and a sequence homologous to the template polynucleotide flanking the variant sequence; (c) combining a building block polynucleotide with a template polynucleotide such that the building block polynucleotide cross-over reassembles with the template polynucleotide to generate polynucleotides comprising homologous gene sequence variations.

10   SLR does not depend on the presence of high levels of homology between polynucleotides to be rearranged. Thus, this method can be used to non-stochastically generate libraries (or sets) of progeny molecules comprised of over 101000 different chimeras. SLR can be used to generate libraries comprised of over 101000 different progeny chimeras. Thus, aspects of the present invention include non-stochastic methods

15 of producing a set of finalized chimeric nucleic acid molecule shaving an overall assembly order that is chosen by design. This method includes the steps of generating by design a plurality of specific nucleic acid building blocks having serviceable mutually compatible ligatable ends, and assembling these nucleic acid building blocks, such that a designed overall assembly order is achieved.

20   The mutually compatible ligatable ends of the nucleic acid building blocks to be assembled are considered to be "serviceable" for this type of ordered assembly if they enable the building blocks to be coupled in predetermined orders. Thus, the overall assembly order in which the nucleic acid building blocks can be coupled is specified by the design of the ligatable ends. If more than one assembly step is to be used, then the

25 overall assembly order in which the nucleic acid building blocks can be coupled is also specified by the sequential order of the assembly step(s). In one aspect, the annealed building pieces are treated with an enzyme, such as a ligase (e.g. T4 DNA ligase), to achieve covalent bonding of the building pieces.

            In one aspect, the design of the oligonucleotide building blocks is obtained

30 by analyzing a set of progenitor nucleic acid sequence templates that serve as a basis for producing a progeny set of finalized chimeric polynucleotides. These parental oligonucleotide templates thus serve as a source of sequence information that aids in the design of the nucleic acid building blocks that are to be mutagenized, e.g., chimerized or shuffled. In one aspect of this method, the sequences of a plurality of parental nucleic

acid templates are aligned in order to select one or more demarcation points. The
demarcation points can be located at an area of homology, and are comprised of one or
more nucleotides. These demarcation points are preferably shared by at least two of the
progenitor templates. The demarcation points can thereby be used to delineate the

5 boundaries of oligonucleotide building blocks to be generated in order to rearrange the
parental polynucleotides. The demarcation points identified and selected in the
progenitor molecules serve as potential chimerization points in the assembly of the final
chimeric progeny molecules. A demarcation point can be an area of homology
(comprised of at least one homologous nucleotide base) shared by at least two parental

10 polynucleotide sequences. Alternatively, a demarcation point can be an area of homology
that is shared by at least half of the parental polynucleotide sequences, or, it can be an
area of homology that is shared by at least two thirds of the parental polynucleotide
sequences. Even more preferably a serviceable demarcation points is an area of
homology that is shared by at least three fourths of the parental polynucleotide sequences,

15 or, it can be shared by at almost all of the parental polynucleotide sequences. In one
aspect, a demarcation point is an area of homology that is shared by all of the parental
polynucleotide sequences.

In one aspect, a ligation reassembly process is performed exhaustively in
order to generate an exhaustive library of progeny chimeric polynucleotides. In other

20 words, all possible ordered combinations of the nucleic acid building blocks are
represented in the set of finalized chimeric nucleic acid molecules. At the same time, in
another aspect, the assembly order (i.e. the order of assembly of each building block in
the 5' to 3 sequence of each finalized chimeric nucleic acid) in each combination is by
design (or non-stochastic) as described above. Because of the non-stochastic nature of

25 this invention, the possibility of unwanted side products is greatly reduced.

In another aspect, the ligation reassembly method is performed
systematically. For example, the method is performed in order to generate a
systematically compartmentalized library of progeny molecules, with compartments that
can be screened systematically, e.g. one by one. In other words this invention provides

30 that, through the selective and judicious use of specific nucleic acid building blocks,
coupled with the selective and judicious use of sequentially stepped assembly reactions, a
design can be achieved where specific sets of progeny products are made in each of
several reaction vessels. This allows a systematic examination and screening procedure
to be performed. Thus, these methods allow a potentially very large number of progeny

molecules to be examined systematically in smaller groups. Because of its ability to perform chimerizations in a manner that is highly flexible yet exhaustive and systematic as well, particularly when there is a low level of homology among the progenitor molecules, these methods provide for the generation of a library (or set) comprised of a large number of progeny molecules. Because of the non-stochastic nature of the instant ligation reassembly invention, the progeny molecules generated preferably comprise a library of finalized chimeric nucleic acid molecules having an overall assembly order that is chosen by design. The saturation mutagenesis and optimized directed evolution methods also can be used to generate different progeny molecular species. It is appreciated that the invention provides freedom of choice and control regarding the selection of demarcation points, the size and number of the nucleic acid building blocks, and the size and design of the couplings. It is appreciated, furthermore, that the requirement for intermolecular homology is highly relaxed for the operability of this invention. In fact, demarcation points can even be chosen in areas of little or no intermolecular homology. For example, because of codon wobble, i.e. the degeneracy of codons, nucleotide substitutions can be introduced into nucleic acid building blocks without altering the amino acid originally encoded in the corresponding progenitor template. Alternatively, a codon can be altered such that the coding for an originally amino acid is altered. This invention provides that such substitutions can be introduced into the nucleic acid building block in order to increase the incidence of intermolecular homologous demarcation points and thus to allow an increased number of couplings to be achieved among the building blocks, which in turn allows a greater number of progeny chimeric molecules to be generated.

In another aspect, the synthetic nature of the step in which the building blocks are generated allows the design and introduction of nucleotides (e.g., one or more nucleotides, which may be, for example, codons or introns or regulatory sequences) that can later be optionally removed in an in vitro process (e.g. by mutagenesis) or in an in vivo process (e.g. by utilizing the gene splicing ability of a host organism). It is appreciated that in many instances the introduction of these nucleotides may also be desirable for many other reasons in addition to the potential benefit of creating a serviceable demarcation point.

In one aspect, a nucleic acid building block is used to introduce an intron. Thus, functional introns are introduced into a man-made gene manufactured according to the methods described herein. The artificially introduced intron(s) can be functional in a

host cells for gene splicing much in the way that naturally-occurring introns serve functionally in gene splicing.

Optimized Directed Evolution System

The methods of the invention also use non-stochastic gene modification system termed "optimized directed evolution system" to generate mutated sequences of the invention, e.g., modified inteins and chimeric polypeptide coding sequences. Optimized directed evolution is directed to the use of repeated cycles of reductive reassortment, recombination and selection that allow for the directed molecular evolution of nucleic acids through recombination. Optimized directed evolution allows generation of a large population of evolved chimeric sequences, wherein the generated population is significantly enriched for sequences that have a predetermined number of crossover events. A crossover event is a point in a chimeric sequence where a shift in sequence occurs from one parental variant to another parental variant. Such a point is normally at the juncture of where oligonucleotides from two parents are ligated together to form a single sequence. This method allows calculation of the correct concentrations of oligonucleotide sequences so that the final chimeric population of sequences is enriched for the chosen number of crossover events. This provides more control over choosing chimeric variants having a predetermined number of crossover events.

In addition, this method provides a convenient means for exploring a tremendous amount of the possible protein variant space in comparison to other systems. Previously, if one generated, for example, 10^{13} chimeric molecules during a reaction, it would be extremely difficult to test such a high number of chimeric variants for a particular activity. Moreover, a significant portion of the progeny population would have a very high number of crossover events which resulted in proteins that were less likely to have increased levels of a particular activity. By using these methods, the population of chimerics molecules can be enriched for those variants that have a particular number of crossover events. Thus, although one can still generate 10^{13} chimeric molecules during a reaction, each of the molecules chosen for further analysis most likely has, for example, only three crossover events. Because the resulting progeny population can be skewed to have a predetermined number of crossover events, the boundaries on the functional variety between the chimeric molecules is reduced. This provides a more manageable number of variables when calculating which oligonucleotide from the original parental polynucleotides might be responsible for affecting a particular trait.

43

One method for creating a chimeric progeny polynucleotide sequence is to create oligonucleotides corresponding to fragments or portions of each parental sequence. Each oligonucleotide preferably includes a unique region of overlap so that mixing the oligonucleotides together results in a new variant that has each oligonucleotide fragment assembled in the correct order. Additional information can also be found, e.g., in USSN 09/332,835; U.S. Patent No. 6,361,974. The number of oligonucleotides generated for each parental variant bears a relationship to the total number of resulting crossovers in the chimeric molecule that is ultimately created. For example, three parental nucleotide sequence variants might be provided to undergo a ligation reaction in order to find a chimeric variant having, for example, greater activity at high temperature. As one example, a set of 50 oligonucleotide sequences can be generated corresponding to each portions of each parental variant. Accordingly, during the ligation reassembly process there could be up to 50 crossover events within each of the chimeric sequences. The probability that each of the generated chimeric polynucleotides will contain oligonucleotides from each parental variant in alternating order is very low. If each oligonucleotide fragment is present in the ligation reaction in the same molar quantity it is likely that in some positions oligonucleotides from the same parental polynucleotide will ligate next to one another and thus not result in a crossover event. If the concentration of each oligonucleotide from each parent is kept constant during any ligation step in this example, there is a 1/3 chance (assuming 3 parents) that an oligonucleotide from the same parental variant will ligate within the chimeric sequence and produce no crossover.

Accordingly, a probability density function (PDF) can be determined to predict the population of crossover events that are likely to occur during each step in a ligation reaction given a set number of parental variants, a number of oligonucleotides corresponding to each variant, and the concentrations of each variant during each step in the ligation reaction. The statistics and mathematics behind determining the PDF is described below. By utilizing these methods, one can calculate such a probability density function, and thus enrich the chimeric progeny population for a predetermined number of crossover events resulting from a particular ligation reaction. Moreover, a target number of crossover events can be predetermined, and the system then programmed to calculate the starting quantities of each parental oligonucleotide during each step in the ligation reaction to result in a probability density function that centers on the predetermined number of crossover events. These methods are directed to the use of repeated cycles of reductive reassortment, recombination and selection that allow for the directed molecular

44

evolution of a nucleic acid encoding a polypeptide through recombination. This system allows generation of a large population of evolved chimeric sequences, wherein the generated population is significantly enriched for sequences that have a predetermined number of crossover events. A crossover event is a point in a chimeric sequence where a shift in sequence occurs from one parental variant to another parental variant. Such a point is normally at the juncture of where oligonucleotides from two parents are ligated together to form a single sequence. The method allows calculation of the correct concentrations of oligonucleotide sequences so that the final chimeric population of sequences is enriched for the chosen number of crossover events. This provides more control over choosing chimeric variants having a predetermined number of crossover events.

In addition, these methods provide a convenient means for exploring a tremendous amount of the possible protein variant space in comparison to other systems. By using the methods described herein, the population of chimerics molecules can be enriched for those variants that have a particular number of crossover events. Thus, although one can still generate 10^{13} chimeric molecules during a reaction, each of the molecules chosen for further analysis most likely has, for example, only three crossover events. Because the resulting progeny population can be skewed to have a predetermined number of crossover events, the boundaries on the functional variety between the chimeric molecules is reduced. This provides a more manageable number of variables when calculating which oligonucleotide from the original parental polynucleotides might be responsible for affecting a particular trait.

In one aspect, the method creates a chimeric progeny polynucleotide sequence by creating oligonucleotides corresponding to fragments or portions of each parental sequence. Each oligonucleotide preferably includes a unique region of overlap so that mixing the oligonucleotides together results in a new variant that has each oligonucleotide fragment assembled in the correct order. See also USSN 09/332,835.

The number of oligonucleotides generated for each parental variant bears a relationship to the total number of resulting crossovers in the chimeric molecule that is ultimately created. For example, three parental nucleotide sequence variants might be provided to undergo a ligation reaction in order to find a chimeric variant having, for example, greater activity at high temperature. As one example, a set of 50 oligonucleotide sequences can be generated corresponding to each portions of each parental variant. Accordingly, during the ligation reassembly process there could be up to

50 crossover events within each of the chimeric sequences. The probability that each of the generated chimeric polynucleotides will contain oligonucleotides from each parental variant in alternating order is very low. If each oligonucleotide fragment is present in the ligation reaction in the same molar quantity it is likely that in some positions oligonucleotides from the same parental polynucleotide will ligate next to one another and thus not result in a crossover event. If the concentration of each oligonucleotide from each parent is kept constant during any ligation step in this example, there is a 1/3 chance (assuming 3 parents) that an oligonucleotide from the same parental variant will ligate within the chimeric sequence and produce no crossover.

Accordingly, a probability density function (PDF) can be determined to predict the population of crossover events that are likely to occur during each step in a ligation reaction given a set number of parental variants, a number of oligonucleotides corresponding to each variant, and the concentrations of each variant during each step in the ligation reaction. The statistics and mathematics behind determining the PDF is described below. One can calculate such a probability density function, and thus enrich the chimeric progeny population for a predetermined number of crossover events resulting from a particular ligation reaction. Moreover, a target number of crossover events can be predetermined, and the system then programmed to calculate the starting quantities of each parental oligonucleotide during each step in the ligation reaction to result in a probability density function that centers on the predetermined number of crossover events.

Iterative Processes

In practicing the invention, these processes can be iteratively repeated. For example a nucleic acid (or, the nucleic acid) responsible for an intein activity is identified, re-isolated, again modified, re-tested for activity. This process can be iteratively repeated until a desired intein phenotype is engineered. An entire biochemical anabolic or catabolic pathway can be engineered into a cell, including proteolytic activity.

Similarly, if it is determined that a particular oligonucleotide has no affect at all on the desired trait, it can be removed as a variable by synthesizing larger parental oligonucleotides that include the sequence to be removed. Since incorporating the sequence within a larger sequence prevents any crossover events, there will no longer be any variation of this sequence in the progeny polynucleotides. This iterative practice of determining which oligonucleotides are most related to the desired trait, and which are

unrelated, allows more efficient exploration all of the possible protein variants that might
be provide a particular trait or activity.

In vivo shuffling

In vivo shuffling of molecules can be used in methods of the invention,
e.g., to develop modified inteins and chimeric polypeptides. In vivo shuffling can be
performed utilizing the natural property of cells to recombine multimers. While
recombination in vivo has provided the major natural route to molecular diversity, genetic
recombination remains a relatively complex process that involves 1) the recognition of
homologies; 2) strand cleavage, strand invasion, and metabolic steps leading to the
production of recombinant chiasma; and finally 3) the resolution of chiasma into discrete
recombined molecules. The formation of the chiasma requires the recognition of
homologous sequences.

In one aspect, the invention provides a method for producing a hybrid
polynucleotide from at least a first polynucleotide and a second polynucleotide. The
invention can be used to produce a hybrid polynucleotide by introducing at least a first
polynucleotide and a second polynucleotide which share at least one region of partial
sequence homology into a suitable host cell. The regions of partial sequence homology
promote processes which result in sequence reorganization producing a hybrid
polynucleotide. The term "hybrid polynucleotide", as used herein, is any nucleotide
sequence which results from the method of the present invention and contains sequence
from at least two original polynucleotide sequences. Such hybrid polynucleotides can
result from intermolecular recombination events which promote sequence integration
between DNA molecules. In addition, such hybrid polynucleotides can result from
intramolecular reductive reassortment processes which utilize repeated sequences to alter
a nucleotide sequence within a DNA molecule.

Producing sequence variants

The methods of the invention introduce one or more mutations into one or
more of nucleic acid segments, e.g., inteins and chimeric sequences of the invention. The
nucleic acids can be altered by any means, including, e.g., random or stochastic methods,
or, non-stochastic, or "directed evolution," methods, as described above.

Mutations can be created using genetic engineering techniques such as site
directed mutagenesis, random chemical mutagenesis, Exonuclease III deletion
procedures, and standard cloning techniques. Alternatively, such variants, fragments,

analogs, or derivatives may be created using chemical synthesis or modification procedures. Other methods of making variants are also familiar to those skilled in the art. These include procedures in which nucleic acid sequences obtained from natural isolates are modified to generate nucleic acids which encode polypeptides having characteristics which enhance their value in industrial or laboratory applications. In such procedures, a large number of variant sequences having one or more nucleotide differences with respect to the sequence obtained from the natural isolate are generated and characterized. These nucleotide differences can result in amino acid changes with respect to the polypeptides encoded by the nucleic acids from the natural isolates.

For example, mutations may be created using error prone PCR. In error prone PCR, PCR is performed under conditions where the copying fidelity of the DNA polymerase is low, such that a high rate of point mutations is obtained along the entire length of the PCR product. Error prone PCR is described, e.g., in Leung, D.W., et al., Technique, 1:11-15, 1989) and Caldwell, R. C. & Joyce G.F., PCR Methods Applic., 2:28-33, 1992. Briefly, in such procedures, nucleic acids to be mutagenized are mixed with PCR primers, reaction buffer, $MgCl_2$, $MnCl_2$, Taq polymerase and an appropriate concentration of dNTPs for achieving a high rate of point mutation along the entire length of the PCR product. For example, the reaction may be performed using 20 fmoles of nucleic acid to be mutagenized, 30 pmole of each PCR primer, a reaction buffer comprising 50 mM KCl, 10 mM Tris HCl (pH 8.3) and 0.01% gelatin, 7 mM $MgCl_2$, 0.5 mM $MnCl_2$, 5 units of Taq polymerase, 0.2 mM dGTP, 0.2 mM dATP, 1 mM dCTP, and 1 mM dTTP. PCR may be performed for 30 cycles of 94°C for 1 min, 45°C for 1 min, and 72°C for 1 min. However, it will be appreciated that these parameters may be varied as appropriate. The mutagenized nucleic acids are cloned into an appropriate vector and the activities of the polypeptides encoded by the mutagenized nucleic acids is evaluated.

Mutations may also be created using oligonucleotide directed mutagenesis to generate site-specific mutations in any cloned DNA of interest. Oligonucleotide mutagenesis is described, e.g., in Reidhaar-Olson (1988) Science 241:53-57. Briefly, in such procedures a plurality of double stranded oligonucleotides bearing one or more mutations to be introduced into the cloned DNA are synthesized and inserted into the cloned DNA to be mutagenized. Clones containing the mutagenized DNA are recovered and the activities of the polypeptides they encode are assessed.

Another method for generating mutations is assembly PCR. Assembly PCR involves the assembly of a PCR product from a mixture of small DNA fragments. A

48

large number of different PCR reactions occur in parallel in the same vial, with the products of one reaction priming the products of another reaction. Assembly PCR is described in, e.g., U.S. Patent No. 5,965,408.

Still another method of generating mutations is sexual PCR mutagenesis. In sexual PCR mutagenesis, forced homologous recombination occurs between DNA molecules of different but highly related DNA sequence in vitro, as a result of random fragmentation of the DNA molecule based on sequence homology, followed by fixation of the crossover by primer extension in a PCR reaction. Sexual PCR mutagenesis is described, e.g., in Stemmer (1994) Proc. Natl. Acad. Sci. USA 91:10747-10751. Briefly, in such procedures a plurality of nucleic acids to be recombined are digested with DNase to generate fragments having an average size of 50-200 nucleotides. Fragments of the desired average size are purified and resuspended in a PCR mixture. PCR is conducted under conditions which facilitate recombination between the nucleic acid fragments. For example, PCR may be performed by resuspending the purified fragments at a concentration of 10-30 ng/:l in a solution of 0.2 mM of each dNTP, 2.2 mM $MgCl_2$, 50 mM KCL, 10 mM Tris HCl, pH 9.0, and 0.1% Triton X-100. 2.5 units of Taq polymerase per 100:l of reaction mixture is added and PCR is performed using the following regime: 94°C for 60 seconds, 94°C for 30 seconds, 50-55°C for 30 seconds, 72°C for 30 seconds (30-45 times) and 72°C for 5 minutes. However, it will be appreciated that these parameters may be varied as appropriate. In some aspects, oligonucleotides may be included in the PCR reactions. In other aspects, the Klenow fragment of DNA polymerase I may be used in a first set of PCR reactions and Taq polymerase may be used in a subsequent set of PCR reactions. Recombinant sequences are isolated and the activities of the polypeptides they encode are assessed.

Mutations may also be created by *in vivo* mutagenesis. In some aspects, random mutations in a sequence of interest are generated by propagating the sequence of interest in a bacterial strain, such as an *E. coli* strain, which carries mutations in one or more of the DNA repair pathways. Such "mutator" strains have a higher random mutation rate than that of a wild-type parent. Propagating the DNA in one of these strains will eventually generate random mutations within the DNA. Mutator strains suitable for use for in vivo mutagenesis are described, e.g., in PCT Publication No. WO 91/16427.

Mutations may also be generated using cassette mutagenesis. In cassette mutagenesis a small region of a double stranded DNA molecule is replaced with a

synthetic oligonucleotide "cassette" that differs from the native sequence. The oligonucleotide often contains completely and/or partially randomized native sequence.

Recursive ensemble mutagenesis may also be used to generate mutations. Recursive ensemble mutagenesis is an algorithm for protein engineering (protein mutagenesis) developed to produce diverse populations of phenotypically related mutants whose members differ in amino acid sequence. This method uses a feedback mechanism to control successive rounds of combinatorial cassette mutagenesis. Recursive ensemble mutagenesis is described, e.g., in Arkin (1992) Proc. Natl. Acad. Sci. USA 89:7811-7815.

In some aspects, mutations are created using exponential ensemble mutagenesis. Exponential ensemble mutagenesis is a process for generating combinatorial libraries with a high percentage of unique and functional mutants, wherein small groups of residues are randomized in parallel to identify, at each altered position, amino acids which lead to functional proteins. Exponential ensemble mutagenesis is described, e.g., in Delegrave (1993) Biotechnology Res. 11:1548-1552. Random and site-directed mutagenesis are described, e.g., in Arnold (1993) Current Opinion in Biotechnology 4:450-455.

In some aspects, mutations are created using shuffling procedures wherein portions of a plurality of nucleic acids which encode distinct polypeptides are fused together to create chimeric nucleic acid sequences which encode chimeric polypeptides as described in, e.g., U.S. Patent Nos. 5,965,408; 5,939,250.

Optimizing codons to achieve high levels of protein expression in host cells

The invention provides methods for modifying intein-encoding nucleic acids (and the nucleic acids encoding the chimeric polypeptides of the invention) to modify codon usage. In one aspect, the invention provides methods for modifying codons in a nucleic acid encoding an intein to increase or decrease its expression in a host cell. The invention also provides nucleic acids encoding an intein modified to increase its expression in a host cell, intein so modified, and methods of making the modified inteins. The method comprises identifying a "non-preferred" or a "less preferred" codon in intein-encoding nucleic acid and replacing one or more of these non-preferred or less preferred codons with a "preferred codon" encoding the same amino acid as the replaced codon and at least one non-preferred or less preferred codon in the nucleic acid has been replaced by a preferred codon encoding the same amino acid. A preferred codon is a codon over-represented in coding sequences in genes in the host cell and a non-preferred or less

preferred codon is a codon under-represented in coding sequences in genes in the host cell.

Host cells for expressing the nucleic acids, expression cassettes and vectors of the invention include bacteria, yeast, fungi, plant cells, insect cells and mammalian cells. Thus, the invention provides methods for optimizing codon usage in all of these cells, codon-altered nucleic acids and polypeptides made by the codon-altered nucleic acids. Exemplary host cells include gram negative bacteria, such as *Escherichia coli* and *Pseudomonas* fluorescens; gram positive bacteria, such as *Streptomyces diversa*, *Lactobacillus gasseri*, *Lactococcus lactis*, *Lactococcus cremoris*, *Bacillus subtilis*. Exemplary host cells also include eukaryotic organisms, e.g., various yeast, such as *Saccharomyces* sp., including *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Pichia pastoris*, and *Kluyveromyces lactis*, *Hansenula polymorpha*, *Aspergillus niger*, and mammalian cells and cell lines and insect cells and cell lines. The invention also includes nucleic acids and polypeptides optimized for expression in these organisms and species.

For example, the codons of a nucleic acid encoding an intein or a chimeric polypeptide of the invention isolated from a bacterial cell are modified such that the nucleic acid is optimally expressed in a bacterial cell different from the bacteria from which the intein was derived, a yeast, a fungi, a plant cell, an insect cell or a mammalian cell. Methods for optimizing codons are well known in the art, see, e.g., U.S. Patent No. 5,795,737; Baca (2000) Int. J. Parasitol. 30:113-118; Hale (1998) Protein Expr. Purif. 12:185-188; Narum (2001) Infect. Immun. 69:7250-7253. See also Narum (2001) Infect. Immun. 69:7250-7253, describing optimizing codons in mouse systems; Outchkourov (2002) Protein Expr. Purif. 24:18-24, describing optimizing codons in yeast; Feng (2000) Biochemistry 39:15399-15409, describing optimizing codons in *E. coli*; Humphreys (2000) Protein Expr. Purif. 20:252-264, describing optimizing codon usage that affects secretion in *E. coli*.

Transgenic non-human animals

The invention provides transgenic non-human animals comprising a nucleic acid (e.g., encoding an intein or a chimeric polypeptide of the invention), a polypeptide, an expression cassette or vector or a transfected or transformed cell of the invention. The invention also provides methods of making and using these transgenic non-human animals.

The transgenic non-human animals can be, e.g., goats, rabbits, sheep, pigs, cows, rats and mice, comprising the nucleic acids of the invention. These animals can be used, e.g., as *in vivo* models to study intein activity, or, as models to screen for agents that change the intein activity *in vivo*. The coding sequences for the polypeptides to be

5 expressed in the transgenic non-human animals can be designed to be constitutive, or, under the control of tissue-specific, developmental-specific or inducible transcriptional regulatory factors. Transgenic non-human animals can be designed and generated using any method known in the art; see, e.g., U.S. Patent Nos. 6,211,428; 6,187,992; 6,156,952; 6,118,044; 6,111,166; 6,107,541; 5,959,171; 5,922,854; 5,892,070; 5,880,327; 5,891,698;

10 5,639,940; 5,573,933; 5,387,742; 5,087,571, describing making and using transformed cells and eggs and transgenic mice, rats, rabbits, sheep, pigs and cows. See also, e.g., Pollock (1999) J. Immunol. Methods 231:147-157, describing the production of recombinant proteins in the milk of transgenic dairy animals; Baguisi (1999) Nat. Biotechnol. 17:456-461, demonstrating the production of transgenic goats. U.S. Patent

15 No. 6,211,428, describes making and using transgenic non-human mammals which express in their brains a nucleic acid construct comprising a DNA sequence. U.S. Patent No. 5,387,742, describes injecting cloned recombinant or synthetic DNA sequences into fertilized mouse eggs, implanting the injected eggs in pseudo-pregnant females, and growing to term transgenic mice whose cells express proteins related to the pathology of

20 Alzheimer's disease. U.S. Patent No. 6,187,992, describes making and using a transgenic mouse whose genome comprises a disruption of the gene encoding amyloid precursor protein (APP).

Polypeptides and peptides

The invention provides chimeric polypeptides comprising at least three

25 domains, wherein the first domain comprises at least one enzyme domain or a binding protein domain, the second domain comprises at least one intein domain and a third domain comprising a detectable moiety domain, at least one intein domain is positioned between at least one enzyme or binding protein and at least one detectable moiety domain, and the intein domain has at least one cleavage or splicing activity. The

30 invention also provides isolated or recombinant polypeptides having a sequence identity to an exemplary sequence of the invention, e.g., SEQ ID NO:2:

Met Glu Lys Thr Glu Lys Asn Glu Leu Val Arg Lys Leu Ile Phe Asn
1   5   10   15

Pro Gln Gly Asp Arg Glu Ala Ser Lys Arg Lys Ile Ile Lys Gly Asn
        20   25   30

5   Pro Thr Asn Ile Phe Glu Leu Asn Glu Ile Lys Tyr Ser Trp Ala Phe
        35   40   45

Asp Leu Tyr Lys Leu Met Gly Phe Thr Asn Phe Trp Ile Pro Glu Glu
    50   55   60

Ile Gln Met Leu Glu Asp Arg Lys Gln Tyr Glu Thr Val Leu Ser Asp
10   65   70   75   80

Tyr Glu Lys Arg Ala Tyr Glu Leu Val Leu Ser Phe Leu Ile Ala Leu
            85   90   95

Asp Ser Phe Gln Val Asp Met Leu Lys Glu Phe Gly Arg Met Ile Thr
            100   105   110

15   Ala Pro Glu Val Glu Met Ala Ile Thr Ala Gln Phe Gln Glu Ser
            115   120   125

Val His Ala Tyr Ser Tyr Gln Phe Ile Leu Glu Ser Val Val Asp Pro
    130   135   140

Val Lys Ala Asp Glu Ile Tyr Asn Tyr Trp Arg Glu Asp Glu Arg Leu
20   145   150   155   160

Leu Glu Arg Asn Lys Val Ile Ala Glu Leu Tyr Asn Glu Phe Ile Arg
            165   170   175

Lys Pro Asn Glu Glu Asn Phe Ile Lys Ala Thr Ile Gly Asn Tyr Ile
            180   185   190

25   Leu Glu Ser Leu Tyr Phe Tyr Ser Gly Phe Ala Phe Phe Tyr Thr Leu
            195   200   205

Gly Arg Gln Gly Lys Met Arg Asn Thr Val Gln Gln Ile Leu Lys Tyr Ile
    210   215   220

Asn Arg Asp Glu Leu Cys Phe Ile Glu Gly Thr Glu Val Leu Thr Lys
30   225   230   235   240

Arg Gly Phe Val Asp Phe Arg Glu Leu Arg Glu Asp Asp Leu Val Ala
            245   250   255

Gln Tyr Asp Ile Glu Thr Gly Glu Ile Ser Trp Thr Lys Pro Tyr Ala
            260   265   270

35   Tyr Val Glu Arg Asp Tyr Glu Gly Ser Met Tyr Arg Leu Lys His Pro

275   280   285

Lys Ser Asn Trp Glu Val Val Ala Thr Glu Gly His Glu Phe Ile Val

290   295   300

Arg Asn Leu Lys Thr Gly Lys Glu Arg Lys Glu Pro Ile Glu Lys Val

305   310   315   320

Lys Leu His Pro Tyr Ser Ala Ile Pro Val Ala Gly Arg Tyr Thr Gly

325   330   335

Glu Val Glu Glu Tyr Asp Leu Trp Glu Leu Val Ser Gly Lys Gly Ile

340   345   350

Thr Leu Lys Thr Arg Ser Ala Val Lys Asn Lys Leu Thr Pro Ile Glu

355   360   365

Lys Leu Leu Ile Val Leu Gln Ala Asp Gly Thr Ile Asp Ser Lys Arg

370   375   380

Asn Gly Lys Phe Thr Gly Phe Gln Gln Leu Lys Phe Phe Phe Ser Lys

385   390   395   400

Tyr Arg Lys Ile Asn Glu Phe Glu Lys Ile Leu Asn Glu Cys Ala Pro

405   410   415

Tyr Gly Ile Lys Trp Lys Lys Tyr Glu Arg Gln Asp Gly Ile Ala Tyr

420   425   430

Thr Val Tyr Tyr Pro Asn Asp Leu Pro Ile Lys Pro Thr Lys Phe Phe

435   440   445

Asp Glu Trp Val Arg Leu Asp Glu Ile Thr Glu Glu Trp Ile Arg Glu

450   455   460

Phe Val Glu Glu Leu Val Lys Trp Asp Gly His Ile Pro Lys Asp Arg

465   470   475   480

Asn Lys Lys Lys Val Tyr Tyr Ser Thr Lys Glu Lys Arg Asn Lys

485   490   495

Asp Phe Val Gln Ala Leu Cys Ala Leu Gly Gly Met Arg Thr Val Val

500   505   510

Ser Arg Glu Arg Asn Pro Lys Ala Lys Asn Pro Val Tyr Arg Ile Trp

515   520   525

Ile Tyr Leu Glu Asp Asp Tyr Ile Asn Thr Gln Thr Met Val Lys Glu

530   535   540

Glu Phe Tyr Tyr Lys Gly Lys Val Tyr Cys Val Ser Val Pro Lys Gly

545   550   555   560

54

Asn Ile Val Val Arg Tyr Lys Asp Ser Val Cys Ile Ala Gly Asn Cys
        565   570   575

His Val Thr Leu Phe Arg Asn Ile Ile Asn Thr Leu Arg Lys Glu Asn
       580   585   590

Pro Glu Leu Phe Thr Pro Glu Ile Glu Lys Trp Ile Val Glu Tyr Phe
     595   600   605

Lys Tyr Ala Val Asn Glu Glu Ile Lys Trp Gly Gln Tyr Val Thr Gln
    610   615   620

Asn Gln Ile Leu Gly Ile Asn Asp Val Leu Ile Glu Arg Tyr Ile Lys
  625   630   635   640

Tyr Leu Gly Asn Leu Arg Ile Thr Gln Ile Gly Phe Asp Pro Ile Tyr
      645   650   655

Pro Glu Val Thr Glu Asn Pro Leu Lys Trp Ile Asp Glu Phe Arg Lys
     660   665   670

Ile Asn Asn Thr Lys Thr Asp Phe Phe Gln Ala Lys Pro Gln Thr Tyr
     675   680   685

Ser Lys Ala Asn Glu Leu Lys Trp
     690   695

Different fluorescent proteins (FP) can be used in the methods and chimeric compositions of the invention. In one aspect, they are compatible with other fluorescent assays, e.g., a variety of colors (green, cyan, red and yellow) can be used. The variety of colors provides flexibility in choosing a FP with spectroscopic activity that does not overlap with that of the enzyme domain or binding protein domain (e.g., the target enzyme).

Polypeptides and peptides can be isolated from natural sources, be synthetic, or be recombinantly generated polypeptides. Peptides and proteins can be recombinantly expressed in vitro or in vivo. The peptides and polypeptides of the invention can be made and isolated using any method known in the art. Polypeptide and peptides of the invention can also be synthesized, whole or in part, using chemical methods well known in the art. See e.g., Caruthers (1980) Nucleic Acids Res. Symp. Ser. 215-223; Horn (1980) Nucleic Acids Res. Symp. Ser. 225-232; Banga, A.K., Therapeutic Peptides and Proteins, Formulation, Processing and Delivery Systems (1995) Technomic Publishing Co., Lancaster, PA. For example, peptide synthesis can be performed using various solid-phase techniques (see e.g., Roberge (1995) Science 269:202; Merrifield (1997) Methods Enzymol. 289:3-13) and automated synthesis may be achieved, e.g.,

using the ABI 431A Peptide Synthesizer (Perkin Elmer) in accordance with the instructions provided by the manufacturer.

The peptides and polypeptides of the invention can also be glycosylated. The glycosylation can be added post-translationally either chemically or by cellular
5   biosynthetic mechanisms, wherein the later incorporates the use of known glycosylation motifs, which can be native to the sequence or can be added as a peptide or added in the nucleic acid coding sequence. The glycosylation can be O-linked or N-linked.

The peptides and polypeptides of the invention, as defined above, include all "mimetic" and "peptidomimetic" forms. The terms "mimetic" and "peptidomimetic"
10 refer to a synthetic chemical compound which has substantially the same structural and/or functional characteristics of the polypeptides of the invention. The mimetic can be either entirely composed of synthetic, non-natural analogues of amino acids, or, is a chimeric molecule of partly natural peptide amino acids and partly non-natural analogs of amino acids. The mimetic can also incorporate any amount of natural amino acid conservative
15 substitutions as long as such substitutions also do not substantially alter the mimetic's structure and/or activity. As with polypeptides of the invention which are conservative variants, routine experimentation will determine whether a mimetic is within the scope of the invention, i.e., that its structure and/or function is not substantially altered. Thus, in one aspect, a mimetic composition is within the scope of the invention if it has an intein
20 activity.

Polypeptide mimetic compositions of the invention can contain any combination of non-natural structural components. In alternative aspect, mimetic compositions of the invention include one or all of the following three structural groups: a) residue linkage groups other than the natural amide bond ("peptide bond") linkages; b)
25 non-natural residues in place of naturally occurring amino acid residues; or c) residues which induce secondary structural mimicry, i.e., to induce or stabilize a secondary structure, e.g., a beta turn, gamma turn, beta sheet, alpha helix conformation, and the like. For example, a polypeptide of the invention can be characterized as a mimetic when all or some of its residues are joined by chemical means other than natural peptide bonds.
30 Individual peptidomimetic residues can be joined by peptide bonds, other chemical bonds or coupling means, such as, e.g., glutaraldehyde, N-hydroxysuccinimide esters, bifunctional maleimides, N,N'-dicyclohexylcarbodiimide (DCC) or N,N'-diisopropylcarbodiimide (DIC). Linking groups that can be an alternative to the traditional amide bond ("peptide bond") linkages include, e.g., ketomethylene (e.g., -

C(=O)-CH$_2$- for -C(=O)-NH-), aminomethylene (CH$_2$-NH), ethylene, olefin (CH=CH),

ether (CH$_2$-O), thioether (CH$_2$-S), tetrazole (CN$_4$-), thiazole, retroamide, thioamide, or

ester (see, e.g., Spatola (1983) in Chemistry and Biochemistry of Amino Acids, Peptides

and Proteins, Vol. 7, pp 267-357, "Peptide Backbone Modifications," Marcell Dekker,

NY).

A polypeptide of the invention can also be characterized as a mimetic by

containing all or some non-natural residues in place of naturally occurring amino acid

residues. Non-natural residues are well described in the scientific and patent literature; a

few exemplary non-natural compositions useful as mimetics of natural amino acid

residues and guidelines are described below. Mimetics of aromatic amino acids can be

generated by replacing by, e.g., D- or L- naphylalanine; D- or L- phenylglycine; D- or L-

2 thieneylalanine; D- or L-1, -2, 3-, or 4- pyreneylalanine; D- or L-3 thieneylalanine; D-

or L-(2-pyridinyl)-alanine; D- or L-(3-pyridinyl)-alanine; D- or L-(2-pyrazinyl)-alanine;

D- or L-(4-isopropyl)-phenylglycine; D-(trifluoromethyl)-phenylglycine; D-

(trifluoromethyl)-phenylalanine; D-p-fluoro-phenylalanine; D- or L-p-biphenyl-

phenylalanine; D- or L-p-methoxy-biphenylphenylalanine; D- or L-2-indole(alkyl)

alanines; and, D- or L-alkylainines, where alkyl can be substituted or unsubstituted

methyl, ethyl, propyl, hexyl, butyl, pentyl, isopropyl, iso-butyl, sec-isotyl, iso-pentyl, or a

non-acidic amino acids. Aromatic rings of a non-natural amino acid include, e.g.,

thiazolyl, thiophenyl, pyrazolyl, benzimidazolyl, naphthyl, furanyl, pyrrolyl, and pyridyl

aromatic rings.

Mimetics of acidic amino acids can be generated by substitution by, e.g.,

non-carboxylate amino acids while maintaining a negative charge; (phosphono)alanine;

sulfated threonine. Carboxyl side groups (e.g., aspartyl or glutamyl) can also be

selectively modified by reaction with carbodiimides (R'-N-C-N-R') such as, e.g., 1-

cyclohexyl-3(2-morpholinyl-(4-ethyl) carbodiimide or 1-ethyl-3(4-azonia- 4,4-

dimetholpentyl) carbodiimide. Aspartyl or glutamyl can also be converted to asparaginyl

and glutaminyl residues by reaction with ammonium ions. Mimetics of basic amino acids

can be generated by substitution with, e.g., (in addition to lysine and arginine) the amino

acids ornithine, citrulline, or (guanidino)-acetic acid, or (guanidino)alkyl-acetic acid,

where alkyl is defined above. Nitrile derivative (e.g., containing the CN-moiety in place

of COOH) can be substituted for asparagine or glutamine. Asparaginyl and glutaminyl

residues can be deaminated to the corresponding aspartyl or glutamyl residues. Arginine

residue mimetics can be generated by reacting arginyl with, e.g., one or more

conventional reagents, including, e.g., phenylglyoxal, 2,3-butanedione, 1,2-cyclo-
hexanedione, or ninhydrin, preferably under alkaline conditions. Tyrosine residue
mimetics can be generated by reacting tyrosyl with, e.g., aromatic diazonium compounds
or tetranitromethane. N-acetylimidizol and tetranitromethane can be used to form O-
acetyl tyrosyl species and 3-nitro derivatives, respectively. Cysteine residue mimetics
can be generated by reacting cysteinyl residues with, e.g., alpha-haloacetates such as 2-
chloroacetic acid or chloroacetamide and corresponding amines; to give carboxymethyl or
carboxyamidomethyl derivatives. Cysteine residue mimetics can also be generated by
reacting cysteinyl residues with, e.g., bromo-trifluoroacetone, alpha-bromo-beta-(5-
imidozoyl) propionic acid; chloroacetyl phosphate, N-alkylmaleimides, 3-nitro-2-pyridyl
disulfide; methyl 2-pyridyl disulfide; p-chloromercuribenzoate; 2-chloromercuri-4
nitrophenol; or, chloro-7-nitrobenzo-oxa-1,3-diazole. Lysine mimetics can be generated
(and amino terminal residues can be altered) by reacting lysinyl with, e.g., succinic or
other carboxylic acid anhydrides. Lysine and other alpha-amino-containing residue
mimetics can also be generated by reaction with imidoesters, such as methyl
picolinimidate, pyridoxal phosphate, pyridoxal, chloroborohydride, trinitro-
benzenesulfonic acid, O-methylisourea, 2,4, pentanedione, and transamidase-catalyzed
reactions with glyoxylate. Mimetics of methionine can be generated by reaction with,
e.g., methionine sulfoxide. Mimetics of proline include, e.g., pipecolic acid, thiazolidine
carboxylic acid, 3- or 4- hydroxy proline, dehydroproline, 3- or 4-methylproline, or 3,3,-
dimethylproline. Histidine residue mimetics can be generated by reacting histidyl with,
e.g., diethylprocarbonate or para-bromophenacyl bromide. Other mimetics include, e.g.,
those generated by hydroxylation of proline and lysine; phosphorylation of the hydroxyl
groups of seryl or threonyl residues; methylation of the alpha-amino groups of lysine,
arginine and histidine; acetylation of the N-terminal amine; methylation of main chain
amide residues or substitution with N-methyl amino acids; or amidation of C-terminal
carboxyl groups.

A residue, e.g., an amino acid, of a polypeptide of the invention can also
be replaced by an amino acid (or peptidomimetic residue) of the opposite chirality. Thus,
any amino acid naturally occurring in the L-configuration (which can also be referred to
as the R or S, depending upon the structure of the chemical entity) can be replaced with
the amino acid of the same chemical structural type or a peptidomimetic, but of the
opposite chirality, referred to as the D- amino acid, but also can be referred to as the R- or
S- form.

The invention also provides methods for modifying the polypeptides of the invention by either natural processes, such as post-translational processing (e.g., phosphorylation, acylation, etc), or by chemical modification techniques, and the resulting modified polypeptides. Modifications can occur anywhere in the polypeptide, including the peptide backbone, the amino acid side-chains and the amino or carboxyl termini. It will be appreciated that the same type of modification may be present in the same or varying degrees at several sites in a given polypeptide. Also a given polypeptide may have many types of modifications. Modifications include acetylation, acylation, ADP-ribosylation, amidation, covalent attachment of flavin, covalent attachment of a heme moiety, covalent attachment of a nucleotide or nucleotide derivative, covalent attachment of a lipid or lipid derivative, covalent attachment of a phosphatidylinositol, cross-linking cyclization, disulfide bond formation, demethylation, formation of covalent cross-links, formation of cysteine, formation of pyroglutamate, formylation, gamma-carboxylation, glycosylation, GPI anchor formation, hydroxylation, iodination, methylation, myristolyation, oxidation, pegylation, proteolytic processing, phosphorylation, prenylation, racemization, selenoylation, sulfation, and transfer-RNA mediated addition of amino acids to protein such as arginylation. See, e.g., Creighton, T.E., Proteins – Structure and Molecular Properties 2nd Ed., W.H. Freeman and Company, New York (1993); Posttranslational Covalent Modification of Proteins, B.C. Johnson, Ed., Academic Press, New York, pp. 1-12 (1983).

Solid-phase chemical peptide synthesis methods can also be used to synthesize the polypeptide or fragments of the invention. Such method have been known in the art since the early 1960's (Merrifield, R. B., J. Am. Chem. Soc., 85:2149-2154, 1963) (See also Stewart, J. M. and Young, J. D., Solid Phase Peptide Synthesis, 2nd Ed., Pierce Chemical Co., Rockford, Ill., pp. 11-12)) and have recently been employed in commercially available laboratory peptide design and synthesis kits (Cambridge Research Biochemicals). Such commercially available laboratory kits have generally utilized the teachings of H. M. Geysen et al, Proc. Natl. Acad. Sci., USA, 81:3998 (1984) and provide for synthesizing peptides upon the tips of a multitude of "rods" or "pins" all of which are connected to a single plate. When such a system is utilized, a plate of rods or pins is inverted and inserted into a second plate of corresponding wells or reservoirs, which contain solutions for attaching or anchoring an appropriate amino acid to the pin's or rod's tips. By repeating such a process step, i.e., inverting and inserting the rod's and pin's tips into appropriate solutions, amino acids are built into desired peptides. In addition, a

number of available FMOC peptide synthesis systems are available. For example, assembly of a polypeptide or fragment can be carried out on a solid support using an Applied Biosystems, Inc. Model 431A™ automated peptide synthesizer. Such equipment provides ready access to the peptides of the invention, either by direct synthesis or by

5   synthesis of a series of fragments that can be coupled using other known techniques.

Inteins

The chimeric compositions of the invention can comprise any known intein, of which there is a wide range of sequence diversity, splicing behavior, and structures. Existing inteins comprise a large family, see, e.g., Perler (2000) Nucleic Acids

10   Res. 28:344-345. The inteins used in the methods of the invention, as part of the chimeric compositions of the invention, including the inteins of the invention, can be recombinantly generated, synthetic, isolated or a combination thereof.

The invention provides an isolated intein of 346 amino acids. This intein was isolated the wild type *Aquifex*. It has a number of rare codons.

15   *Methods for modifying the expression of inteins*

The invention provides methods for modifying the function and expression of inteins, including the inteins of the invention. The methods can comprise any protocol for modifying nucleic acid and/or polypeptide sequences and/or structures. For example, the methods of the invention comprise the use of GSSM™, or Gene Site Saturation

20   Mutagenesis™. GSSM is a method of systematically generating every possible amino acid mutation at every residue of a particular gene. In one aspect, intein mutagenesis focuses on intein regions responsible for splicing.

When fused to a target gene, this intein dramatically reduces gene expression. This may be because of an intein's large size or sub-optimal codon usage. In

25   one aspect, the intein-modifying procedures develop an intein that has little or no affect on gene expression. In one aspect, an intein of the invention has little or no affect on gene expression when fused to a target gene.

In one aspect, the intein-modifying procedures identify the minimal intein regions necessary for optimal intein expression without affecting splicing activity. For

30   many inteins, the relatively small domains at the N- and C-terminus of the intein are required for splicing activity; see, e.g., ., Chong (1997) J. of Biol. Chem. 272:15587-15590; Telenti (1997) J. of Bacteriol. 179:6378-6382; Lew (1998) J. of Biol. Chemistry 273:15887-15890. Thus, with only routine screening it can be determined which regions

of the *Aquifex* intein of the invention can be removed without affecting its behavior. In one aspect, the invention provides intein variants where sequences not responsible for activity, e.g., splicing activity, are removed. In one aspect, the deleted intein sequences are replaced with another sequence, e.g., a protein coding sequence, such as a detectable protein, e.g., a fluorescent protein (FP). In one aspect, the invention provides an intein whose non-splicing sequences have been replaced with a detectable protein (e.g., a fluorescent protein). In one aspect, this sequence is placed in an expression system, e.g., a vector. In one aspect, this chimeric sequence is expressed and used to normalize gene expression.

The direct fusion of an intein and a detectable protein, e.g., an FP, may lead to steric constraints that prevent either intein splicing or detection, e.g., FP fluorescence. In one aspect, linkers are added to the N- and/or C- terminal of the detectable protein, e.g., FP. In one aspect, coding sequences for linkers are added to the N- and/or C- terminal of the coding sequence of the detectable protein. In one aspect, at the same time that the intein deletions are screened, a variety of linkers can be added to the N- and/or C- terminal of the detectable protein, e.g., FP; see Figure 9. In one aspect, at the same time that the intein deletions are screened, coding sequences for a variety of linkers can be added to the coding sequences of the N- and/or C- terminal of the detectable protein, e.g., FP.

In one aspect, the minimal intein constructs are screened for function (e.g., splicing, enzymatic activity) in the pEKI1 construct (pEKint1, see Figure 2). In one aspect, this only provides kanamycin resistance if the intein is able to splice. In one aspect, using the kanamycin selection for intein activity and fluorescent detection of FP function, a large number of intein/linker combinations are screened to find a minimal intein that retains splicing and FP function (see Figure 9). In one aspect, owing to its smaller size, this "mini-intein" expresses at higher levels when fused to a target protein.

In one aspect, an intein used in the compositions and/or methods of the invention, including the intein of the invention, is modified by optimizing its codon usage for high expression in various screening hosts, e.g., yeast, bacteria and the like. Codon optimization of a minimal intein may be easier than of a full-length intein. In one aspect, these modifications generate an active fluorescent intein that can be fused to a target gene (NPT II, beta-gal) without significantly reducing its expression.

In one aspect, the invention provides methods for modifying inteins, particularly intein splicing regions, e.g., intein splicing regions in the chimeric

compositions of the invention. Critical regions for intein splicing have been extensively characterized, see, e.g., Chong (1997) J. of Biol. Chem. 272:15587-15590; Telenti (1997) J. of Bacteriol. 179:6378-6382; Chong (1996) J. of Biol. Chem. 271:22159-22168; Derbyshire (1997) Proc. Natl. Acad. Sci. USA 94:11466-11471; Perler (1997) Nucleic Acids Res. 25:1087-1093. Thus, in one aspect, an intein used in the compositions and/or methods of the invention, including the intein of the invention, comprises only intein sequence(s) critical for splicing.

In one aspect, an advantage of an intein-detectable moiety (e.g., FP) chimeric construct of the invention over a traditional gene fusion to a detectable moiety (e.g., FP) is that the resulting splicing that liberates the protein of interest (encoded by a gene of interest, GOI) from the intein-detectable moiety (e.g., FP) reporter section of the chimeric composition. This can be important because in many cases the fusion of even a small affinity tag to either end of a gene or polypeptide can significantly alter gene and/or polypeptide expression, nucleic acid and/or protein stability and/or activity. In this case, the splicing or cleavage of a fusion tag is necessary for liberating an unaltered protein of interest/ gene of interest. If the gene:intein-detectable moiety fusion does not splice efficiently, in some aspects, the advantage of using this intein technology may be diminished or lost. Therefore, in one aspect, intein constructs with good splicing efficiencies are provided. The invention also provides methods for making and screening for intein constructs with good splicing efficiencies.

In one aspect, the invention provides methods for intein modification comprising evolving conditional mutants to effect *in vitro* splicing for protein purification, see, e.g., Telenti (1997) J. of Bacteriol. 179:6378-6382; Wood (1999) Nature Biotechnology 17: 889-892; Southworth (2000) EMBO Journal 19:5019-5026. In one aspect, intein modification comprises *in vivo* or *in vitro* splicing behavior. Mutations that decreased or increased *in vivo* activity were screened.

In one aspect, the invention provides methods for determining important intein domains for catalytic activity or other modification of intein behavior. In alternative aspects, inteins are evolved to have temperature, pH and DTT dependence and/or *in vitro* and/or *in vivo* cleavage activity, see, e.g., Perler (2000) Nucleic Acids Res. 28:344-345.

The pEKI1 (pEKInt1) construct will only provide kanamycin resistance if the intein splices. Many of the mutations generated by GSSM or other techniques will most likely inactivate intein activity. In one aspect, using the pEKI1 construct, a large

library of intein mutants is screened for clones able to splice the intein and, subsequently, grow in the presence of kanamycin, see Figure 10. In one aspect, positive clones are characterized in a quantitative assay for kanamycin activity to determine which clones splice more efficiently. Quantitative assays for kanamycin activity are known in the art,

5   see, e.g., Henderson (1991) Analytical Biochemistry 194:64-68. In one aspect, the invention provides a method comprising a combination of a powerful antibiotic selection and a quantitative assay. This can enable the rapid evaluation of large numbers of intein mutants to evolve a more efficiently splicing intein.

In alternative aspects, the intein used in the methods and the compositions

10 of the invention have a splicing efficiency of over about 70%, 75%, 80%, 85%, 90%, 95%, 97%, 99%, or more. In one aspect, the amount of unspliced fusion protein (e.g., KanR:Intein-FP) is measured by protein gel. In one aspect, the amount of unspliced fusion protein is less than about 1%, 5%, 10%, 15%, 20% or more of the spliced protein (e.g., KanR). In one aspect, estimation of protein levels is by SDS-PAGE and/or by mass

15 spectrometer (MS). MS can quickly and accurately determine the exact size and amount of proteins present in a sample. If large numbers of samples require quantitation of splicing efficiency, mass spectrometry can be used.

Modifying inteins for cleavage rather than splicing

In one aspect, the intein of the invention and the inteins used in the

20 chimeric compositions and polypeptides of the invention have only splicing or only cleaving activity (including substantially splicing or substantially cleaving activity). In one aspect, the intein of the invention and the inteins used in the chimeric compositions and polypeptides of the invention exhibit cleavage rather than splicing when fused to the C-terminus of a target gene. In one aspect, the intein of the invention exhibits cleavage

25 rather than splicing when fused to the C-terminus of a target gene. This cleavage behavior may be used to broaden the applicability of the intein vectors of the invention. In one aspect, see Figure 11, the intein is fused between the target gene as an "N-extein" and a His6 affinity tag as a "C-extein." In one aspect, after splicing, the target gene is rejoined to the affinity tag. If cleavage rather than splicing occurs, the target gene is

30 liberated with its native C-terminus and the affinity tag remains fused to the intein. In many cases, it may be necessary to retain the native C-terminus of the target gene for proper protein function.

The pEKI2 (pEKInt2) construct (see Fig. 2) will only confer kanamycin resistance if the intein cleaves rather than splices. If the intein were to splice, the His6 tag would remain attached to the NPT II enzyme, which would result in its inactivation. In the pEKI2 (pEKInt2) construct the intein cleaves. Thus, in one aspect pEKI2 (pEKInt2) construct is used to select for intein mutants that undergo significant cleavage.

In one aspect, gene evolution technologies are combined with the selection screening for intein cleavage to generate intein constructs that undergo efficient cleavage rather than splicing. In alternative aspects, greater than about 70%, 75%, 80%, 85%, 90%, 95%, 97%, 98%, 99%, or more of the expressed protein is cleaved as assayed by protein gel and/or mass spectrometry.

Modifying inteins to be insensitive to −1 AA position

In one aspect, the intein of the invention and the inteins used in the chimeric compositions and polypeptides of the invention function efficiently regardless of the target gene to which they are joined. Also, in one aspect, the intein of the invention and the inteins used in the chimeric compositions and polypeptides of the invention function generate greater than about 70%, 75%, 80%, 85%, 90%, 95%, 97%, 98%, 99%, or more splicing and/or cleavage with a variety of target genes.

Some proteins, e.g., enzymes, may adversely affect intein activity. The amino acid residue located at the N-terminus of an intein (i.e., at the −1 position) can significantly affect splicing activity. In one aspect, intein constructs with different target genes are screened with a wide range of amino acids at the −1 position. In one aspect, the intein is evolved to decrease its sensitivity to the sequence located at the C-terminus of the intein.

Screening Methodologies and "On-line" Monitoring Devices

In practicing the methods of the invention, a variety of apparatus and methodologies can be used to in conjunction with the polypeptides and nucleic acids of the invention, e.g., to screen polypeptides for intein activity, to screen compounds as
5   potential modulators, e.g., activators or inhibitors, of an intein activity, for antibodies that bind to a polypeptide of the invention, for nucleic acids that hybridize to a nucleic acid of the invention, to screen for cells expressing a polypeptide of the invention and the like. In one aspect, invention provides activity screens using the chimeric compositions of the invention to normalize gene expression, wherein the chimeric compositions of the
10 invention can splice out (cleave out) an enzyme domain or a binding protein domain with about 70%, 75%, 80%, 85%, 90%, 95%, 97%, 98%, 99%, or more efficiency with a variety of target genes.

Capillary Arrays

Capillary arrays, such as the GIGAMATRIX™, Diversa Corporation, San
15 Diego, CA, can be used to in the methods of the invention. Nucleic acids or polypeptides of the invention can be immobilized to or applied to an array, including capillary arrays. Arrays can be used to screen for or monitor libraries of compositions (e.g., small molecules, antibodies, nucleic acids, etc.) for their ability to bind to or modulate the activity of a nucleic acid or a polypeptide of the invention. Capillary arrays provide
20 another system for holding and screening samples. For example, a sample screening apparatus can include a plurality of capillaries formed into an array of adjacent capillaries, wherein each capillary comprises at least one wall defining a lumen for retaining a sample. The apparatus can further include interstitial material disposed between adjacent capillaries in the array, and one or more reference indicia formed within
25 of the interstitial material. A capillary for screening a sample, wherein the capillary is adapted for being bound in an array of capillaries, can include a first wall defining a lumen for retaining the sample, and a second wall formed of a filtering material, for filtering excitation energy provided to the lumen to excite the sample.

A polypeptide or nucleic acid, e.g., a ligand, can be introduced into a first
30 component into at least a portion of a capillary of a capillary array. Each capillary of the capillary array can comprise at least one wall defining a lumen for retaining the first component. An air bubble can be introduced into the capillary behind the first component. A second component can be introduced into the capillary, wherein the

second component is separated from the first component by the air bubble. A sample of interest can be introduced as a first liquid labeled with a detectable particle into a capillary of a capillary array, wherein each capillary of the capillary array comprises at least one wall defining a lumen for retaining the first liquid and the detectable particle, and wherein the at least one wall is coated with a binding material for binding the detectable particle to the at least one wall. The method can further include removing the first liquid from the capillary tube, wherein the bound detectable particle is maintained within the capillary, and introducing a second liquid into the capillary tube.

The capillary array can include a plurality of individual capillaries comprising at least one outer wall defining a lumen. The outer wall of the capillary can be one or more walls fused together. Similarly, the wall can define a lumen that is cylindrical, square, hexagonal or any other geometric shape so long as the walls form a lumen for retention of a liquid or sample. The capillaries of the capillary array can be held together in close proximity to form a planar structure. The capillaries can be bound together, by being fused (e.g., where the capillaries are made of glass), glued, bonded, or clamped side-by-side. The capillary array can be formed of any number of individual capillaries, for example, a range from 100 to 4,000,000 capillaries. A capillary array can form a micro titer plate having about 100,000 or more individual capillaries bound together.

Arrays, or "Biochips"

Nucleic acids or polypeptides of the invention can be immobilized to or applied to an array. Arrays can be used to screen for or monitor libraries of compositions (e.g., small molecules, antibodies, nucleic acids, etc.) for their ability to bind to or modulate the activity of a nucleic acid or a polypeptide of the invention. For example, in one aspect of the invention, a monitored parameter is transcript expression of an intein gene. One or more, or, all the transcripts of a cell can be measured by hybridization of a sample comprising transcripts of the cell, or, nucleic acids representative of or complementary to transcripts of a cell, by hybridization to immobilized nucleic acids on an array, or "biochip." By using an "array" of nucleic acids on a microchip, some or all of the transcripts of a cell can be simultaneously quantified. Alternatively, arrays comprising genomic nucleic acid can also be used to determine the genotype of a newly engineered strain made by the methods of the invention. Polypeptide arrays" can also be used to simultaneously quantify a plurality of proteins. The present invention can be practiced with any known "array," also referred to as a "microarray" or "nucleic acid

array" or "polypeptide array" or "antibody array" or "biochip," or variation thereof.
Arrays are generically a plurality of "spots" or "target elements," each target element
comprising a defined amount of one or more biological molecules, e.g., oligonucleotides,
immobilized onto a defined area of a substrate surface for specific binding to a sample
5   molecule, e.g., mRNA transcripts.

      In practicing the methods of the invention, any known array and/or method
of making and using arrays can be incorporated in whole or in part, or variations thereof,
as described, for example, in U.S. Patent Nos. 6,277,628; 6,277,489; 6,261,776;
6,258,606; 6,054,270; 6,048,695; 6,045,996; 6,022,963; 6,013,440; 5,965,452; 5,959,098;
10   5,856,174; 5,830,645; 5,770,456; 5,632,957; 5,556,752; 5,143,854; 5,807,522; 5,800,992;
5,744,305; 5,700,637; 5,556,752; 5,434,049; see also, e.g., WO 99/51773; WO 99/09217;
WO 97/46313; WO 96/17958; see also, e.g., Johnston (1998) Curr. Biol. 8:R171-R174;
Schummer (1997) Biotechniques 23:1087-1092; Kern (1997) Biotechniques 23:120-124;
Solinas-Toldo (1997) Genes, Chromosomes & Cancer 20:399-407; Bowtell (1999)
15   Nature Genetics Supp. 21:25-32. See also published U.S. patent applications Nos.
20010018642; 20010019827; 20010016322; 20010014449; 20010014448; 20010012537;
20010008765.

Adapting existing screening methods to include normalization

      Almost every quantitative screen can be improved by normalizing activity
20   to gene expression. Thus, in one aspect, the compositions and methods of the invention
are used in quantitative screens to normalize gene expression. The compositions and
methods of the invention can be used with any quantitative screen. For any given
quantitation method (e.g., fluorescence, absorbance, HPLC, MS, bioactivity, etc.), a
relatively simple step is added to assay the fluorescence of the chimeric compositions of
25   the invention (e.g., the intein-FP).

      For enzyme assays, any screening platform can be used. For example, the
methods of the invention can include microtiter plates, e.g., with 96, 384, or 1536 wells
each. In an exemplary enzyme assay, clones expressing β-galactosidase activity are
detected by adding the chromogenic substrate 2-nitrophenyl β-D-galactopyranoside
30   (ONPG) to bacteria grown in microtiter plates. In one aspect, the invention uses
sensitive, cell-permeable, fluorogenic substrates like β-D-galactopyranoside (e.g., by
Molecular Probes #R-1159). In one aspect, the cells can be grown in the presence of the
substrate. In one aspect, the invention provides enzyme activity assays involving multiple

67

steps, e.g., multiple liquid transfer steps, cell lysis step(s), and characterization(s) of
reaction products by HPLC, MS or other means.

In one aspect, the methods of the invention are practiced in wells, e.g.,
microtiter plate wells. In one aspect, smaller well volumes are used to increase
throughput and decrease reagent costs. However, smaller well volumes may also increase
variability due to growth, expression, and liquid transfer differences. In one aspect,
GigaMatrix™ is used in the methods of the invention. In one aspect, GigaMatrix™ is
used for fluorescent detection from samples in very small wells, for example, as small as
250 nL. A GigaMatrix™ plate can contain 100,000 wells in the same footprint as a
standard microtiter plate. This screening format may offer advantages in screening
throughput and reagent use. The GigaMatrix™ system has been used to screen libraries
for a number of enzyme classes using different fluorogenic substrates.

In one aspect, the invention provides very high throughput assays using,
e.g., a Fluorescence Activated Cell Sorter (FACS) to assay enzyme activity in a single cell
format. In one aspect, a FACS instrument measures multiple fluorescent signals from
single cells at a rate greater than 10^7 cells/min. In one aspect, the invention uses a
substrate that is cell-permeable, non-toxic and/or non-fluorescent. In one aspect, when
modified by the appropriate enzyme, the substrate becomes fluorescent and remains
inside the cell instead of diffusing out. Several classes of enzymes can be used in the
chimeric compositions of the invention and used in methods of the invention comprising
a FACS assay (e.g., to assay enzyme activity in a single cell format).

Any assay can be used in the methods of the invention for gene expression
normalization. In invention provides assays with sample sizes decreasing from millions
of cells in a microtiter plate to a single cell in a FACS assay. As the sample sizes
decrease, the variations in growth and expression become more significant. In a larger
sample size, these variations average out in a larger population of cells. Normalization
using the compositions and methods of the invention can correct for variations and enable
accurate, quantitative measurements of enzyme activity.

In one aspect, the normalized assays of the invention and/or compositions
of the invention are used to develop enzymes for the production of high value chiral
compounds. These assays can use, e.g., racemases, nitrilases and epoxide hydrolases. In
one aspect, the normalized assays of the invention and/or compositions of the invention
are used to develop enzymes for the biocatalytic production of high value pharmaceutical
compounds. In one aspect, the normalized assays of the invention and/or compositions of

the invention are used to develop enzymes, e.g., racemases, hydrolases and nitrilases, that exhibit a variety of different properties and activity profiles, including varied expression, stability, pH and temperature optimums, substrate specificity and enantioselectivity. In one aspect, the normalized assays of the invention and/or compositions of the invention

5   are used in the enantioselective production of high value amino acids. High throughput spectroscopic methods can be used. The exemplary method described in Figure 12 can be used to quantitate the enantioselective conversion of amino acids using secondary detection enzymes.

Figure 12 illustrates an exemplary method for the quantitation of

10   enantioselective nitrilase conversion. The D-amino acid oxidase (D-AAO) and L-amino acid oxidase (L-AAO) reactions can be set up in parallel on microtiter plates, such that one is subjected to treatment with the L-processing enzyme and the other is treated with the D-processing enzyme. The amount of L-amino acid generated is then detected by converting it to the alpha-keto acid. This oxidation is accompanied by a concomitant

15   release of ammonia and hydrogen peroxide. In one aspect, the H_2O_2 generated is then detected, e.g., using a highly specific and sensitive assay, such as using horseradish peroxidase conducted in the presence of a chromogenic hydrogen donor, such as phenol:4-amino-antipyrine whose oxidation yields a red chromophore. Enantioselectivity can be calculated by comparison of the parallel assays.

20   The enantioselectivity of the nitrilase reaction will be assayed in parallel using a D- specific or L-specific amino acid oxidase. Preliminary research has shown that levels of D- and L- amino acids can be measured by splitting the nitrilase reaction mixture into two and assaying using the secondary enzyme reporter system. The enantioselective detection of chiral amino acids using this assay was successfully

25   demonstrated using D- and L- alanine, phenylalanine, and several unnatural amino acids. This assay can be used to quantitatively screen libraries of enzyme mutants, e.g., those generated by GSSM and GeneReassembly technologies. Screening libraries for increased enantioselectivity using this method may not be sensitive to differences in growth and enzyme expression because the calculation is based on the ratio of L- to D- amino acids

30   and is independent of the total amount of enzyme conversion. However, normalizing activity levels using the intein-FP vectors may allow mutants to be selected based on both their enantioselectivity and their specific activity. Furthermore, measuring the intein-FP fluorescence in the secondary reactions may correct for liquid transfer errors introduced when the original nitrilase reaction is split into two.

The invention provides a single step assay using the normalization methods and compositions of the invention. In one aspect, the nitrilase reaction and the secondary amino acid oxidase reaction are performed concurrently. This can eliminate the need to split the nitrilase reaction into two separate assays. The parallel assays can be performed on duplicate samples grown separately. By eliminating the liquid transfer steps, this can reduce the time and effort of this assay and enable the reduction of reaction size to 1536 well plates or smaller. The utilization of compositions of the invention (e.g., a chimeric intein-FP) and normalization methods of the invention can reduce screening efforts and cost, increase sensitivity and productivity, and enable evolution of more effective nitrilases for the production of high value chiral amino acids.

In alternative aspects assays, e.g., racemases, nitrilase and hydrolase assays, are adapted to capillary array systems, e.g., the GigaMatrix™ screening platforms. In one aspect, GigaMatrix™ wells can be filled with growth media containing an aminonitrile substrate, amino acid oxidase secondary enzyme and a fluorescent hydrogen donor to detect H_2O_2 production. E. coli harboring the library of nitrilase mutants to be screened can be diluted such that an average of one cell per well will be added to each well. As each mutant clone grows and expresses the nitrilase, the production of a D- or L- amino acid can generate a fluorescent signal due to the coupled reaction with the amino acid oxidase. In one aspect, the production of both the L- and D- amino acids is directly measured. In one aspect, the methods of the invention provide a very rapid primary screen to detect mutants that produce a certain enantiomer. Positive clones can be recovered from the wells and characterized in more detail using a secondary assay. Variations in growth and expression can be even more dramatic in small volumes of the GigaMatrix™ wells compared to the larger well of a microtiter plate. In this aspect, the addition of gene expression normalization methods of the invention greatly enhances the sensitivity and quantitative of these assays.

In one aspect, the bacterial clones grow in the reaction media. In one aspect, use of GigaMatrix™ varies depending on a number of factors, including substrate/product, permeability/toxicity and fluorescence sensitivity. Figure 13 describes an exemplary nitrilase activity assay using the fluorogenic reagent dihydroxyphenoxazine. This method has been successfully used with several different aminonitriles, demonstrating that these substrates are sufficiently non-toxic and cell permeable to allow sensitive detection.

          In another aspect, the chimeric compositions and methods of the invention are used in enzyme assays, e.g., racemase, nitrilase or epoxide hydrolase assays, for the production of chiral compositions, e.g., epoxides and diols, e.g., for use as pharmaceutical intermediates. In one aspect, the methods comprise spectroscopic and fluorescence based measurements of enzyme activities, e.g., epoxide hydrolase activity. These assays can utilize synthesized substrate analogs to generate a chromogenic or fluorescent product, as illustrated in Figure 14.

          Figure 14 schematically illustrates exemplary epoxide hydrolase screens. Figure 14A illustrates an exemplary fluorogenic epoxide assays using a fluorogenic epoxide substrate. Epoxide hydrolase activity can generate a fluorescent signal when this synthetic substrate is cleaved. Figure B illustrates an exemplary chromogenic epoxide assay. Epoxide hydrolase activity can reduce the color change when this assay is performed.

          To validate this assay, chromogenic and fluorogenic epoxide hydrolase substrates were synthesized at Diversa. Using these substrates in the assay, the activity of a putative epoxide hydrolase was verified. As in the nitrilase assay, the fluorescent detection of epoxide hydrolase activity could be adapted to high density microtiter plates or the GigaMatrix format. Using the intein-FP technologies, these assays will be more quantitative, enabling the development of additional epoxide hydrolase enzymes that will produce specific high value compounds.

          In addition to spectroscopic assays in, e.g., microtiter plates, GigaMatrix™, and FACS, activity in the methods of the invention is measured by a variety of biological and analytical methods. Any assay providing a quantitative value for activity can be normalized to protein expression using the methods of the invention. Intein-detectable moiety (e.g., intein-FP) fluorescence can be measured in a variety of formats.

          In alternative aspects, the normalized assays, including both the nitrilase and epoxide hydrolase assays of the invention, are used in high throughput screening. In one aspect, the methods of the invention provide increased data quality and sensitivity and lower background "noise" from variations in growth, expression, and pipetting. In one aspect, the additional information provided by the normalization methods of the invention is used to compare enzymes based on their specific activity rather than whole cell activity.

Screening libraries for modified proteins using intein normalized data

In one aspect, the invention provides methods for screening libraries for modified enzymes or binding proteins (e.g., improved or other altered activity) using intein normalized data generated by the methods of the invention. In one aspect, a

5 specific target for this project is nitrilases, e.g., those for the production of intermediates for ACE inhibitors. In one aspect, the goal is to evolve an appropriate enzyme to produce key intermediates, e.g., 4-phenyl-2-amino-butanoic acid, for manufacturing these drugs.

The invention also provides methods for developing a process for using this enzyme to manufacture the 4-phenyl-2-amino-butanoic acid intermediate. This

10 includes developing a production process for 4-phenyl-2-amino-butanoic acid. In alternative aspects, the processes include various enzyme properties, including variations in enantioselectivity, specific activity, stability, expression, etc. The methods of the invention can incorporate any nucleic acid or polypeptide modification schemes, including GSSM and/or GeneReassembly technologies described above. In one aspect,

15 mutant libraries are constructed and screened for the desired enzyme properties. Hits from this screening effort can be characterized in more detail and can serve as a template for a second round of evolution to further modify the enzyme's characteristics. This iterative cycle of mutagenesis and screening can be repeated until a desired protein, e.g., enzyme or binding protein, has been developed.

20 The intein technology of the invention provides an efficient means to express genes in a one-to-one ratio with a detectable label, e.g., a fluorescent protein. Thus, the amount of protein produced by a single cell or a population of cells can be calculated from a fluorescent measurement of the sample. The methods of the invention can increase the sensitivity and productivity of various screening platforms and enable the

25 development of novel screening methods. The methods of the invention for normalizing gene expression in living cells have many applications in academic and industrial areas.

The invention provides novel screens for the development of enzymes for the production of high value chiral molecules. These enzymes can catalyze the selective (enantioselective) conversion of molecules with the chirality. The enzymes generated by

30 this technology can be used in production processes to meet demand for chiral molecules discussed in this document.

The methods of the invention can be used to discover, develop, manufacture and employ enzymes for the production of chiral intermediates and active pharmaceutical ingredients for pharmaceuticals, such as chiral building blocks, enzyme libraries and bulk enzymes. The methods of the invention can be used for nitrilase and epoxide hydrolase biotransformation. The exemplary enzyme targets for use with the methods of the invention are set forth in the following table:

Nitrilase Targets		Epoxide Hydrolase Targets	
ACE inhibitors intermediate		**Multi-outlet Chiral Synthons**	
• Cilazapril	• Lisinopril	• Glycidol	
• Temocapril	• Quinapril	• Epichlorohydrin	
• Benazepril	• Enalapril	• Propylene oxide	
• Moexipril	• Ramipril	• Methyl glycidate	
• Imidapril	• Trandolapril	• Ethyl glycidate	
• Fosinopril		• Styrene oxide	
		• 3-Chlorostyrene oxide	
		• Phenyl glycidyl ether	
		• Benzyl glycidyl ether	
Profens		**Oxazolidinone Antibiotics**	
• Ketoprofen	• Loxoprofen	• Linezolid	
• Ibuprofen	• Naproxen		
• Suprofen	• Dexketoprofen		
• Zaltoprofen	• Fenoprofen		
• Pirprofen	• Flunoxaprofen		
• Suprofen	• Pirprofen		
		Antifungals	
Others			
• Lipitor intermediate	• L-carnitine	• Posaconazole (phase III clinicals, GSK)	
• Paclitaxel	• R-mandelic acid	• Ravuconazole (phase II clinicals, BMS)	
• Docetaxel	• D-phenylglycine	• Ketoconazole	
• D-hydroxyphenylglycine	• L-tert leucine	• Fluconazole	
• Ubenimex	• R-4-Fluoromandelic acid	• Griseofulvin	
• R-2-Chloromandelic acid		• Terbinafine	
• R-3-Chloromandelic acid	• R-2-Methylmandelic acid	• Flucytosine	
• S-Phenyllactic acid		• Itraconazole	
• S-Azetidine-2-carboxylic acid	• L-Homophenylalanine	• Amphotericin B	
• D-lactic acid		• Voriconazole	
		Antivirals	
		• Amprenavir	• Oseltamivir
		• Telinavir	• Ganciclovir
		• Famciclovir	• Saquinavir

• Entecavir	• Ribavirin
• Zidovudine	• Zanamivir
• Ritonavir	• Pirodavir
• Indinavir	• Lamivudine
• Nelfinavir	• Famciclovir
• Vidarabine	• Didanosine
• Zalcitabine	• Edoxudine
• Lobucavir	• Cidofovir
	• Tenofovir
	• Abacavir
Beta Blockers	
• Metoprolol	• Celiprolol
• Atenolol	• Carvedilol
• Bisoprolol	• Timolol
• Betaxolol	• Nadolol
• Acebutolol	• Penbutolol
• Carteolol	• Pindolol
• Esmolol	• Propanolol
Beta Agonists	
• Formoterol	
• Terbutaline	
• Salmeterol	
• Albuterol	
• Metaproteronol	
• Salbutamol	
•	
Steroid Derivatives	
• Fluticasone	
• Triamincinolone	

Gene reassembly to generate intein libraries

The chimeric compositions of the invention can comprise any known intein, of which there is a wide range of sequence diversity, splicing behavior, and structures. These sequences can be modified by any technique or methodology, e.g., using gene reassembly techniques, e.g., GeneReassembly™, to recombine elements of the different inteins to generate diverse libraries. Thus, in one aspect, the invention provides a method comprising GeneReassembly™ to recombine domains from a family of genes. The method of the invention combines domains from different inteins to create inteins and chimeric proteins with novel properties.

An exemplary method of the invention for making libraries of inteins using GeneReassembly™ is set forth in Figure 8. In one aspect, these libraries are screened for mutants with different (e.g., increased) splicing and/or cleavage efficiency.

The libraries can be cloned into an cloning vehicle, e.g., a vector, to facilitate the screening and/or to recombine elements of the different inteins.

When genes or domains are recombined using GeneReassembly™, the number of potential combinations can be large. For example, if the 4 intein splicing motifs from 10 different intein genes or domains were recombined, there would be 410 = 1,048,576 possible combinations. Adding more parent genes or recombining subdomains of the splicing motifs would generate even larger numbers. Routine screening can identify active recombined mutants. In one aspect, antibiotic selection is used in the routine intein activity screening assays. This greatly reduces the number of clones that need to be evaluated, making the assay more quantitative. If the splicing efficiency of only a relatively small number of clones (1 to 1000) needs to be quantitatively evaluated, an antibiotic immunoactivity assay can be used, e.g., a Kanamycin immunoactivity assay as described, e.g., in Henderson (1991) Analytical Biochem. 194:64-68. In one aspect, this assay involves immunoaffinity immobilization of an NPTII protein (the protein conferring kanamycin resistance, also called KanR) in the sample. In one aspect, it is followed by a secondary enzyme reaction to measure NPTII activity.

The invention provides methods comprising the combination of a selection screen and a high throughput activity screen. For example, an exemplary methods comprise a growth selection for nitrilase activity and a quantitative high throughput nitrilase assay. In one aspect, instead of the NPTII gene, a nitrilase is used a construct for the selection and activity screen. In one aspect, a mutant library is selected in a bacterial host, e.g., an *E. coli* host, that requires an active nitrilase to grow on a minimal medium supplemented with a nitrile as the sole nitrogen source. Only clones containing an active intein will have nitrilase activity and thus the ability to grow on this medium. In one aspect, these selected clones can be then be assayed in a high throughput activity screen to measure nitrilase activity using a fluorescent substrate as an activity indicator. In one aspect, this screen has been used to screen up to 20,000 clones or more per day and detect increases in activity as low as 10%.

Antibodies and Antibody-based screening methods

The invention provides isolated or recombinant antibodies that specifically bind to an intein of the invention. These antibodies can be used to isolate, identify or quantify the inteins of the invention or related polypeptides. These antibodies can be

used to isolate other polypeptides within the scope the invention or other related inteins. The antibodies can be designed to bind to a splice or cleavage site of an intein.

The antibodies can be used in immunoprecipitation, staining, immunoaffinity columns, and the like. If desired, nucleic acid sequences encoding for specific antigens can be generated by immunization followed by isolation of polypeptide or nucleic acid, amplification or cloning and immobilization of polypeptide onto an array of the invention. Alternatively, the methods of the invention can be used to modify the structure of an antibody produced by a cell to be modified, e.g., an antibody's affinity can be increased or decreased. Furthermore, the ability to make or modify antibodies can be a phenotype engineered into a cell by the methods of the invention.

Methods of immunization, producing and isolating antibodies (polyclonal and monoclonal) are known to those of skill in the art and described in the scientific and patent literature, see, e.g., Coligan, CURRENT PROTOCOLS IN IMMUNOLOGY, Wiley/Greene, NY (1991); Stites (eds.) BASIC AND CLINICAL IMMUNOLOGY (7th ed.) Lange Medical Publications, Los Altos, CA ("Stites"); Goding, MONOCLONAL ANTIBODIES: PRINCIPLES AND PRACTICE (2d ed.) Academic Press, New York, NY (1986); Kohler (1975) Nature 256:495; Harlow (1988) ANTIBODIES, A LABORATORY MANUAL, Cold Spring Harbor Publications, New York. Antibodies also can be generated in vitro, e.g., using recombinant antibody binding site expressing phage display libraries, in addition to the traditional in vivo methods using animals. See, e.g., Hoogenboom (1997) Trends Biotechnol. 15:62-70; Katz (1997) Annu. Rev. Biophys. Biomol. Struct. 26:27-45.

Polypeptides or peptides can be used to generate antibodies which bind specifically to the polypeptides, e.g., the inteins, of the invention. The resulting antibodies may be used in immunoaffinity chromatography procedures to isolate or purify the polypeptide or to determine whether the polypeptide is present in a biological sample. In such procedures, a protein preparation, such as an extract, or a biological sample is contacted with an antibody capable of specifically binding to one of the polypeptides of the invention.

In immunoaffinity procedures, the antibody is attached to a solid support, such as a bead or other column matrix. The protein preparation is placed in contact with the antibody under conditions in which the antibody specifically binds to one of the polypeptides of the invention. After a wash to remove non-specifically bound proteins, the specifically bound polypeptides are eluted.

The ability of proteins in a biological sample to bind to the antibody may be determined using any of a variety of procedures familiar to those skilled in the art. For example, binding may be determined by labeling the antibody with a detectable label such as a fluorescent agent, an enzymatic label, or a radioisotope. Alternatively, binding of the antibody to the sample may be detected using a secondary antibody having such a detectable label thereon. Particular assays include ELISA assays, sandwich assays, radioimmunoassays, and Western Blots.

Polyclonal antibodies generated against the polypeptides of the invention can be obtained by direct injection of the polypeptides into an animal or by administering the polypeptides to a non-human animal. The antibody so obtained will then bind the polypeptide itself. In this manner, even a sequence encoding only a fragment of the polypeptide can be used to generate antibodies which may bind to the whole native polypeptide. Such antibodies can then be used to isolate the polypeptide from cells expressing that polypeptide.

For preparation of monoclonal antibodies, any technique which provides antibodies produced by continuous cell line cultures can be used. Examples include the hybridoma technique, the trioma technique, the human B-cell hybridoma technique, and the EBV-hybridoma technique (see, e.g., Cole (1985) in Monoclonal Antibodies and Cancer Therapy, Alan R. Liss, Inc., pp. 77-96).

Techniques described for the production of single chain antibodies (see, e.g., U.S. Patent No. 4,946,778) can be adapted to produce single chain antibodies to the polypeptides of the invention. Alternatively, transgenic mice may be used to express humanized antibodies to these polypeptides or fragments thereof.

Antibodies generated against the polypeptides of the invention may be used in screening for similar polypeptides (e.g., inteins) from other organisms and samples. In such techniques, polypeptides from the organism are contacted with the antibody and those polypeptides which specifically bind the antibody are detected. Any of the procedures described above may be used to detect antibody binding.

Kits

The invention provides kits comprising the compositions, e.g., nucleic acids, expression cassettes, vectors, cells, transgenic seeds or plants or plant parts, polypeptides (e.g., inteins) and/or antibodies of the invention. The kits also can contain

instructional material teaching the methodologies and industrial uses of the invention, as described herein.

Measuring Metabolic Parameters

The methods of the invention for normalizing gene expression in assays can be used in whole cell evolution, or whole cell engineering, of a cell to develop a new cell strain having a new phenotype, e.g., a new or modified enzyme activity, a modified intein activity. In one aspect, the genetic composition of the cell is modified. The genetic composition can be modified by addition to the cell of a nucleic acid of the invention. To detect the new phenotype, at least one metabolic parameter of a modified cell is monitored in the cell in a "real time" or "on-line" time frame. In one aspect, a plurality of cells, such as a cell culture, is monitored in "real time" or "on-line." In one aspect, a plurality of metabolic parameters is monitored in "real time" or "on-line." Metabolic parameters can be monitored using the inteins of the invention.

Metabolic flux analysis (MFA) is based on a known biochemistry framework. A linearly independent metabolic matrix is constructed based on the law of mass conservation and on the pseudo-steady state hypothesis (PSSH) on the intracellular metabolites. In practicing the methods of the invention, metabolic networks are established, including the:

   • identity of all pathway substrates, products and intermediary metabolites
   • identity of all the chemical reactions interconverting the pathway metabolites, the stoichiometry of the pathway reactions,
   • identity of all the enzymes catalyzing the reactions, the enzyme reaction kinetics,
   • the regulatory interactions between pathway components, e.g. allosteric interactions, enzyme-enzyme interactions etc,
   • intracellular compartmentalization of enzymes or any other supramolecular organization of the enzymes, and,
   • the presence of any concentration gradients of metabolites, enzymes or effector molecules or diffusion barriers to their movement.

Once the metabolic network for a given strain is built, mathematic presentation by matrix notion can be introduced to estimate the intracellular metabolic fluxes if the on-line metabolome data is available. Metabolic phenotype relies on the changes of the whole metabolic network within a cell. Metabolic phenotype relies on the change of pathway utilization with respect to environmental conditions, genetic

regulation, developmental state and the genotype, etc. In one aspect of the methods of the invention, after the on-line MFA calculation, the dynamic behavior of the cells, their phenotype and other properties are analyzed by investigating the pathway utilization. For example, if the glucose supply is increased and the oxygen decreased during the yeast fermentation, the utilization of respiratory pathways will be reduced and/or stopped, and the utilization of the fermentative pathways will dominate. Control of physiological state of cell cultures will become possible after the pathway analysis. The methods of the invention can help determine how to manipulate the fermentation by determining how to change the substrate supply, temperature, use of inducers, etc. to control the physiological state of cells to move along desirable direction. In practicing the methods of the invention, the MFA results can also be compared with transcriptome and proteome data to design experiments and protocols for metabolic engineering or gene shuffling, etc.

In practicing the methods of the invention, any modified or new phenotype can be conferred and detected, including new or improved characteristics in the cell. Any aspect of metabolism or growth can be monitored.

Detectable moieties

The invention provides a chimeric protein comprising at least three domains, wherein one domain comprises a detectable moiety domain. The detectable moiety can be any detectable composition, including fluorescent, bioluminescent, chemiluminescent or radioactive moieties. Bioluminescence includes all bioluminescence, fluorescence or chemiluminescence or other photon detectable systems.

Bioluminescent and chemiluminescent polypeptides used in the compositions and methods of the invention include all known polypeptides known to be bioluminescent or chemiluminescent, or, acting as enzymes on a specific substrate (reagent), can generate (by their enzymatic action) a bioluminescent or chemiluminescent molecule. They include, e.g., isolated and recombinant luciferases, aequorin, obelin, mnemiopsin, berovin and variations thereof and combinations thereof, as discussed in detail, below. In some aspects, the bioluminescent or chemiluminescent are enzymes that act on a substrate that reacts with the reagent *in situ* to generate a molecule that can be imaged. The substrate can be administered before, at the same time (e.g., in the same formulation), or after administration of the chimeric polypeptide (including the enzyme).

In alternative aspects, these polypeptides include, e.g., luciferase, aequorin, halistaurin, phialidin, obelin, mnemiopsin or berovin, or, equivalent

photoproteins, and combinations thereof. The compositions and methods of the invention also include recombinant forms of these polypeptides as recombinant chimeric or "fusion" proteins, including chimeric nucleic acids and constructs encoding them. Methods of making recombinant forms of these polypeptides are well known in the art,

5   e.g., luciferase reporter plasmids are described, e.g., by Everett (1999) J. Steroid Biochem. Mol. Biol. 70:197-201. Sala-Newby (1998) Immunology 93:601-609, described used of a recombinant cytosolic fusion protein of firefly luciferase and aequorin (luciferase-aequorin). The Ca^{2+}-activated photoprotein obelin is described by, e.g., Dormer (1978) Biochim. Biophys. Acta 538:87-105; and, recombinant obelin is described

10 by, e.g., Illarionov (2000) Methods Enzymol. 305:223-249. The photoprotein mnemiopsin is described by, e.g., Anctil (1984) Biochem J. 221:269-272. The monomeric $Ca2+$-binding protein aequorin is described by, e.g., Kurose (1989) Proc. Natl. Acad. Sci. USA 86:80-84; Shimomura (1995) Biochem. Biophys. Res. Commun. 211:359-363. The aequorin-type photoproteins halistaurin and phialidin are described by,

15 e.g., Shimomura (1985) Biochem J. 228:745-749. Ward (1975) Proc. Natl. Acad. Sci USA 72:2530-2534, describes the purification of mnemiopsin, aequorin and berovin. The recombinant bioluminescent or chemiluminescent chimeric polypeptides of the invention can be made by any method, see, e.g., U.S. Patent No. 6,087,476, that describes making recombinant, chimeric luminescent proteins. U.S. Patent Nos. 6,143,50; 6,074,859;

20 6,074,859, 5,229,285, describe making recombinant luminescent proteins. The bioluminescent or chemiluminescent activity of the chimeric recombinant polypeptides of the invention can be assayed, e.g., using assays described in, e.g., U.S. Patent Nos. 6,132,983; 6,087,476; 6,060,261; 5,866,348; 5,094,939; 5,744,320. Various photoproteins that can be used in compositions of the invention are described in, e.g.,

25 U.S. Patent Nos. 5,648,218; 5,360,728; 5,098,828.

           The compositions and methods of the invention include use of any device capable of detecting bioluminescence, fluorescence or chemiluminescence or other photon detection systems. The methods of the invention can be practiced using any photon detection device, or variation or equivalent thereof, or in conjunction with any

30 known photon detection methodology, including visual imaging. An exemplary photodetector device is an intensified charge-coupled device (ICCD) camera coupled to an image processor. See, e.g., U.S. Patent No. 5,650,135. Photon detection devices are manufactured by, e.g., Xenogen (Alameda, CA) (the Xenogen IVIS™ imaging system); or, Hamamatsu Corp., Bridgewater, NJ.

Another imaging system used for compositions and methods of the invention can be a proximal charge-coupled device (CCD) detection/imaging. Due to its inherent versatility, it can also accommodate chemiluminescence, fluorescent and radioisotope target molecule detection, high throughput, and high sensitivity. This detection/imaging apparatus can include a lensless imaging array comprising a plurality of solid state imaging devices, such as an array of CCDs, photoconductor-on-MOS arrays, photoconductor-on-CMOS arrays, charge injection devices (CIDs), photoconductor on thin-film transistor arrays, amorphous silicon sensors, photodiode arrays, or the like.

The compositions and methods of the invention incorporate in whole or in part designs of detection devices as described, e.g., in U.S. Patent Nos. 6,197,503; 6,197,498; 6,150,147; 6,083,763; 6,066,448; 6,045,996; 6,025,601; 5,599,695; 5,981,956; 5,698,089; 5,578,832; 5,632,957.

The compositions and methods of the invention incorporate and use any detectable label. Exemplary labels include, e.g., ^{32}P, ^{35}S, ^{3}H, ^{14}C, ^{125}I, ^{131}I; fluorescent dyes (e.g., Cy5™, Cy3™, FITC, rhodamine, lanthanide phosphors, Texas red), electron-dense reagents (e.g. gold), enzymes, e.g., as commonly used in an ELISA (e.g., horseradish peroxidase, beta-galactosidase, luciferase, alkaline phosphatase), colorimetric labels (e.g. colloidal gold), magnetic labels (e.g. Dynabeads™), biotin, dioxigenin, or haptens and proteins for which antisera or monoclonal antibodies are available. The label can be directly incorporated into the nucleic acid or other target compound to be detected, or it can be attached to a probe or antibody that hybridizes or binds to the target. A peptide can be made detectable by incorporating (e.g., into a nucleoside base) predetermined polypeptide epitopes recognized by a secondary reporter (e.g., leucine zipper pair sequences, binding sites for secondary antibodies, transcriptional activator polypeptide, metal binding domains, epitope tags). Label can be attached by spacer arms of various lengths to reduce potential steric hindrance or impact on other useful or desired properties (the domains of the chimeric compositions can also be separated by various spacers). See, e.g., Mansfield (1995) Mol Cell Probes 9:145-156.

Cyanine and related dyes, such as merocyanine, styryl and oxonol dyes, are particularly strongly light-absorbing and highly luminescent, see, e.g., U.S. Patent Nos. 4,337,063; 4,404,289; 6,048,982. Cy3™ and Cy5™ can be used together; both are fluorescent cyanine dyes produced by Amersham Life Sciences (Arlington Heights, IL).

In the compositions and methods of the invention, labeling with a detectable composition (labeling with a detectable moiety) also can include a nucleic acid attached to another biological molecule, such as a nucleic acid, e.g., a nucleic acid in the form of a stem-loop structure as a "molecular beacon" or an "aptamer beacon."
Molecular beacons as detectable moieties are well known in the art; for example, Sokol (1998) Proc. Natl. Acad. Sci. USA 95:11538-11543, synthesized "molecular beacon" reporter oligodeoxynucleotides with matched fluorescent donor and acceptor chromophores on their 5' and 3' ends. In the absence of a complementary nucleic acid strand, the molecular beacon remains in a stem-loop conformation where fluorescence resonance energy transfer prevents signal emission. On hybridization with a complementary sequence, the stem-loop structure opens increasing the physical distance between the donor and acceptor moieties thereby reducing fluorescence resonance energy transfer and allowing a detectable signal to be emitted when the beacon is excited by light of the appropriate wavelength. See also, e.g., Antony (2001) Biochemistry 40:9387-9395, describing a molecular beacon comprised of a G-rich 18-mer triplex forming oligodeoxyribonucleotide. See also U.S. Patent Nos. 6,277,581 and 6,235,504.

Aptamer beacons are similar to molecular beacons; see, e.g., Hamaguchi (2001) Anal. Biochem. 294:126-131; Poddar (2001) Mol. Cell. Probes 15:161-167; Kaboev (2000) Nucleic Acids Res. 28:E94. Aptamer beacons can adopt two or more conformations, one of which allows ligand binding. A fluorescence-quenching pair is used to report changes in conformation induced by ligand binding. See also, e.g., Yamamoto (2000) Genes Cells 5:389-396; Smirnov (2000) Biochemistry 39:1462-1468.

In addition to methods for labeling nucleic acids with fluorescent dyes, methods for the simultaneous detection of multiple fluorophores are well known in the art, see, e.g., U.S. Patent Nos. 5,539,517; 6,049,380; 6,054,279; 6,055,325. For example a spectrograph can image an emission spectrum onto a two-dimensional array of light detectors; a full spectrally resolved image of the array is thus obtained. Photophysics of the fluorophore, e.g., fluorescence quantum yield and photodestruction yield, and the sensitivity of the detector are read time parameters for an oligonucleotide array.

Examples

Example 1: Development and assessment of chimeric proteins and methods of the
invention.

The following example describes the development and assessment of
exemplary chimeric proteins and methods of the invention.

The feasibility of using an intein containing a fluorescent protein (FP) to
monitor gene expression *in vivo* was demonstrated. Several exemplary intein expression
vectors were constructed and used to achieve this goal. In various aspects, the vector can
comprise a fluorescent protein (FP) inserted into an intein and fused downstream of a
gene of interest (GOI), see Figure 1, a schematic describing the strategy for developing an
exemplary intein expression vector. Any intein can be used in the compositions and
methods of the invention, and many are commercially available, e.g., fluorescent proteins
from Clontech Inc. In these exemplary chimeric proteins, inteins from the vacuolar
ATPase subunit of *Saccharomyces cerevisiae* (Sce VMA) and the Green Fluorescent
Protein (GFP) intein were used. Additionally, a new intein from the bacterium *Aquifex
aeolicus* was found, as discussed below.

The intein's unique autocatalytic properties were exploited in order to
achieve a stoichiometric expression of the GOI and the FP. This technology permitted the
normalization of an enzyme's activity to its associated FP fluorescence. This intein
cleaves at its N-terminus when placed at the C-terminal end of a protein. The strategy
used to demonstrate the feasibility of this exemplary intein expression vector is
summarized in Figure 1.

The intein from the ribonucleoside diphosphate reductase (nrdF) gene of
the *Aquifex aeolicus* genome was determined by sequence homology to contain an intein
encoding 346 amino acid residues. Its ability to catalyze its own excision from its native
protein precursor and from within the context of a foreign protein was not previously
demonstrated. FP's from Clontech in a variety of different colors were also used; they
behave in a similar manner to GFP in terms of their bacterial expression and tolerance of
N- or C-terminal fusions. The neomycin phosphotransferase gene (NPT II) (hereafter
denoted as KanR) provides kanamycin resistance and was used as a marker for
preliminary screening of the intein fusions.

The entire intein-containing *nrdF* gene was PCR amplified from *Aquifex
aeolicus* (SEQ ID NO:1) and overexpressed in *E. coli*. Protein gel analysis clearly

showed efficient splicing of the *Aquifex* intein from its native gene, confirming its intein activity.

When most inteins splice from their natural protein precursor, they leave behind a cysteine at the site of insertion. Thus, it is possible to insert the intein sequence at any cysteine in a target gene. Constructs were made with the *Aquifex* intein cloned into either cysteine33 or at the C-terminus of the KanR gene and subcloned into the pET21b (Novagen) vector behind the T7 promoter to generate plasmids pEKI1 (pEKint1) and pEKI2 (pEKint2), respectively; see Figure 2, a diagram of the constructs used to generate exemplary intein expression vector.

The KanR gene does not normally have a cysteine at its C-terminus, so one was added for the construction of pEKI2. The KanR:Intein fusions were also subcloned as fusions to a hexahistidine (His$_6$) tag for affinity purification and Western detection of the spliced products. As a control, KanR was subcloned in the same vector with its native C-terminus, an additional cysteine at the C-terminus or a His$_6$ affinity tag to generate pEK, pEKCys, and pEKH respectively. *E. coli* carrying each plasmid was tested for growth in the presence of kanamycin at 25 µg/ml (see Fig. 2). The pEK plasmid conferred kanamycin resistance, but pEKH and pEKCys did not. The addition of a single cysteine residue at the C-terminus of the KanR gene destroyed its ability to confer kanamycin resistance. As a consequence, it was not expected that any of the His$_6$ tagged constructs would provide kanamycin resistance, even if the intein spliced and regenerated the full length KanR protein.

The pEKI1 (pEKInt1) plasmid was able to confer kanamycin resistance, confirming the splicing activity of the *Aquifex* intein in a foreign context. The pEKI2H (pEKInt2H) plasmid was also able to confer kanamycin resistance. Combined with the fact that a His$_6$ tag destroys KanR function, this strongly suggested that the kanamycin resistance of the pEKI2H was a result of a cleavage event at the N-terminus of the intein instead of splicing. In this case, the His$_6$ tag remained fused to the intein and the KanR protein retained its native C-terminus; see Figure 3, a schematic diagram of splicing versus cleavage of the *Aquifex aeolicus* intein.

Following overexpression in *E. coli*, whole cell extracts were prepared and analyzed by SDS-PAGE, as illustrated in see Figure 4. In Figure 4, overexpression of intein fusions in BL21Star(DE3)pLysS was analyzed: pEK = kanamycin alone; pEKH = kanamycin with His6 tag; pEKI1(pEKint1) = kanamycin with intein at Cys 33; pEKI1H = kanamycin plus intein at Cys 33 with His6 tag; pEKI2 = kanamycin with C-terminal

intein; pEKI2H = kanamycin with C-terminal intein and His6 tag. U = uninduced; I = induced. K = kanamycin; KH = kanamycin with His6 tag; KI1H = kanamycin with intein and His6 tag; I1 = spliced intein; KI2H = kanamycin with intein and His6 tag; I2H = intein with His6 tag.

The predicted molecular sizes of KanR (the native KanR protein) and KanRH (His6 tagged KanR protein) are 29 and 30 kDa, respectively (SEQ ID NO:1 encoding SEQ ID NO:2). In the pEKI1H sample, a 30 kDa was visible, which likely corresponds to KanRH. This strongly suggests that the intein spliced and KanR containing the His6 tag was regenerated. In contrast, a 29 kDa protein corresponding to KanR was detected in the pEKI2H sample providing solid evidence that the intein cleaved at its N-terminus. In this case, the His6 tag remained fused to the intein and KanR was unaltered. These results paralleled those from the kanamycin resistance screening.

Similar constructs were created using the Sce VMA intein as a comparison. Using protein gel analysis, see Figure 5, and the kanamycin resistance selection as an assay for splicing, it was shown that the VMA intein also splices correctly when inserted into cysteine33 of the KanR gene. However, the VMA construct analogous to the pKanInt2H was not able to confer kanamycin resistance, indicating that the VMA intein does not undergo cleavage under these conditions. These results demonstrate that this *in vivo* cleavage activity is a property of the *Aquifex* intein. Figure 5 illustrates the overexpression of VMA intein in BL21(DE3)RIL. Lane 1 = pEK induced; 2 = pEKH induced; 3 = pEK33V uninduced; 4 = pEK33V induced; 5 = pEK33VH uninduced; 6 = pEK33VH induced; KVMAH = kanamycin plus VMA intein and His6 tag; VMA = intein alone; KH = kanamycin with His6 tag.

Three fluorescent proteins (FP) were individually inserted into the intein constructs: a red shifted variant of green fluorescent protein (EGFP) (excitation 488 nm; emission 507 nm) from plasmid pIJ8641, see, e.g. Sun (1999) Microbiology 145:2221-2227; the cyan fluorescent protein (excitation 453 nm; emission 486 nm) from *Anemonia majano* in vector pAmCyan (Clontech); and a variant of the wild-type green fluorescent protein (excitation 496 nm; emission 506 nm) from *Zoanthus* sp. in vector pZsGreen (Clontech). From sequence comparison with other inteins, it was predicted that a region of approximately 500 bp should not be required for splicing or cleavage of the *Aquifex* intein. Vectors were generated by replacing this region of the intein in the pEKI2H plasmid with each FP.

Overexpression of the intein-FP constructs in *E. coli* resulted in high-level expression of each fusion; however, neither splicing nor cleavage was apparent. Consequently, the fusion did not confer resistance to kanamycin. Whole cell fluorescence was assayed but was not detected, which may be a result of inactivation of EGFP as an
5   internal fusion. In one aspect, flexible linkers are added to the chimeric polypeptides of the invention. In one aspect, the detectable moiety, e.g., fluorescent protein, is fused the C-terminus of the intein, see Figure 6, which illustrates exemplary chimeric structures that retain intein function and fluorescence. Figure 6A: Flexible linkers are introduced between the intein and FP domains to relax steric interference between the domains.
10   Figure B: Fusion of the GFP C-terminal to the intein. After translation, the intein will cleave, releasing the GOI without any C-terminal additions.

The Green and Cyan FP's from Clontech were inserted into the VMA intein at residue 204 or replacing residues 204-383. These residues were shown to be dispensable for intein activity if replaced by a small flexible linker; see, e.g., Chong
15   (1997) J. of Biol. Chem. 272:15587-15590. The lacZ gene from *E. coli* and two other β-galactosidase genes were fused in place of KanR to the N-terminus of the VMA and VMA-FP constructs. These β-gal genes were chosen because they represent a range of specific activities and expression levels. When overexpressed in *E. coli*, the constructs with Green or Cyan FP had no detectable fluorescence in either of the intein
20   configurations, which may be due to the fusion location. As determined by SDS-PAGE, none of the constructs appeared to splice, although fusion proteins of expected sizes were produced. All of these constructs were positive for β-galactosidase activity using several different substrates to detect activity. Figure 7 illustrates β-gal fusions to VMA-FP inteins. All combinations of three different β-gal genes, two intein configurations
25   (insertion at 204 or deletion of 204-384), and two FP's (Clontech Green and Cyan) were generated and overexpressed in *E. coli*.

The exemplary *Aquifex* intein chosen for this project proved to be difficult to express and had significantly reduced splicing activity when inserted into a foreign context. When constructs were made by inserting the different FP genes into the *Aquifex*
30   or VMA inteins, neither intein activity nor FP fluorescence were observed. Manipulation of the VMA and other inteins has shown that a flexible linker between the N- and C-terminal intein motifs is required (Chong (1997) J. of Biol. Chem. 272:15587-15590). The configurations in which the FP's were inserted may have caused structural interference leading to a nonfunctional intein. This interference may also explain the lack

of FP fluorescence. The addition of flexible linkers between the intein and FP sequences may relax the conformation allowing the intein and FP domains to function normally. Alternatively, in one aspect, the FP is fused to the C-terminus of the intein, as illustrated in Figure 6. The cleavage activity of the *Aquifex* intein produces the desired protein.

Although fluorescence was not detected from these fusions constructed in these experiments, it is likely that the location of FP insertion affects its function. Several examples from the literature have successfully inserted GFP into a gene and retained fluorescence, see, e.g., Biondi (1998) Nucleic Acids Res. 26:4946-4952. It may be that the choice of insertion of the FP's into the intein constructs described above does not allow for correct FP folding and function.

The invention provides chimeric compositions (e.g., fusion proteins) comprising linker insertions between their domains. The intein domains may require proper folding for efficient splicing or cleavage to occur. The reacting groups may have to align precisely. In a study by Chong (1997) J. of Biol. Chem. 272:15587-15590, a portion of the intein encoding an endonuclease was deleted. The resulting mini-intein failed to exhibit splicing. However, two flexible peptide linkers were inserted into the site that was deleted, which resulted in efficient splicing *in vivo*. Another example of a linker insertion to improve the splicing of a mini-intein was demonstrated in the Mxe GyrA intein from *Mycobacterium xenopi*, see, e.g., Telenti (1997) J. of Bacteriol. 179:6378-6382. Any linker, including those described by Chong (1997) or Telenti (1997), can be used in chimeric compositions of the invention. Routine screening can be done on known inteins to generate optimal mini-inteins for particular purposes.

In one aspect, chimeric compositions of the invention further comprise a domain for affinity purification, such as a chitin binding domain, a His6 tag, any epitope for Ab affinity ligation, and the like. The chimeric compositions of the invention can be a mini-intein vector comprising a domain for affinity purification. In one aspect, the polypeptide can be a modified intein fused to a target protein at its N-terminus and a chitin binding domain-GFP fusion at its C-terminus, as described by Zhang (2001) Gene 275:241-252. This permits monitoring of soluble protein expression by following GFP fluorescence combined with an option for affinity purification. In Zhang (2001), cleavage was only observed *in vitro*. However, in Zhang (2001), an intein fused in frame to GFP resulted in a functional intein albeit *in vitro* and a functional fluorescence. The Zhang (2001) data demonstrates that the fluorescence of a GFP fluorescence-intein after cleavage (e.g., from a chimeric composition as illustrated in Figure 6B) can be monitored.

Exemplary chimeric compositions comprising inteins can undergo cleavage instead of splicing and can be used in normalization technology. In one aspect, inteins are used that enable the expression of gene products having their native C-termini. In one alternative, a small affinity tag is ligated at the C-terminus.

In one aspect, antibiotic selections are used for both splicing and cleavage events. These selections can be used to develop additional, including more efficient, inteins in Phase II.

In one aspect, a pH sensitive GFP is incorporated into a chimeric composition of the invention. It can be used to detect proton production by an enzyme, e.g., a hydrolytic enzyme. Thus, in one aspect, a chimeric composition comprises a pH sensitive GFP and a hydrolytic enzyme (and an intein).

A number of embodiments of the invention have been described. Nevertheless, it will be understood that various modifications may be made without departing from the spirit and scope of the invention. Other embodiments are within the scope of the following claims.

WHAT IS CLAIMED IS:

1.   A chimeric protein comprising at least three domains, wherein the first domain comprises at least one enzyme domain or a binding protein domain, the second domain comprises at least one intein domain and a third domain comprising a detectable moiety domain, at least one intein domain is positioned between at least one enzyme or binding protein and at least one detectable moiety domain, and the intein domain has at least one cleavage or splicing activity.

2.   The chimeric protein of claim 1, wherein the enzyme is a nitrilase, a kinase, a racemase or a hydrolase.

3.   The chimeric protein of claim 2, wherein the hydrolase is an epoxide hydrolase, a phosphatase, a lipase or a protease.

4.   The chimeric protein of claim 1, wherein the binding protein is an antibody or a receptor.

5.   The chimeric protein of claim 1, wherein the intein comprises a polypeptide encoded by a nucleic acid sequence as set forth in SEQ ID NO:1.

6.   The chimeric protein of claim 1, wherein the intein comprises a sequence as set forth in SEQ ID NO:2.

7.   The chimeric protein of claim 1, wherein the detectable moiety domain comprises a detectable peptide or polypeptide.

8.   The chimeric protein of claim 7, wherein the detectable peptide or a polypeptide is a fluorescent peptide or polypeptide.

9.   The chimeric protein of claim 7, wherein the detectable peptide or a polypeptide is a bioluminescent or a chemiluminescent peptide or polypeptide.

10.   The chimeric polypeptide of claim 9, wherein the bioluminescent or chemiluminescent polypeptide comprises a green fluorescent protein (GFP), an aequorin, an obelin, a mnemiopsin or a berovin.

11.   The chimeric protein of claim 1, wherein the detectable moiety domain comprises an enzyme that generates a detectable signal.

12.   The chimeric protein of claim 11, wherein the enzyme that generates a detectable signal comprises an alpha-galactosidase, a chloramphenicol acetyltransferase or a kinase.

13.   The chimeric protein of claim 1, wherein the detectable moiety domain comprises a radioactive isotope.

14.   The chimeric protein of claim 1, wherein the chimeric protein is a recombinant fusion protein.

15.   The chimeric protein of claim 1, wherein the intein domain splicing activity results in cleavage of the enzyme domain from the intein domain and detectable domain.

16.   The chimeric protein of claim 1, wherein the intein domain splicing activity results in cleavage of the enzyme domain from the intein domain and detectable domain and cleavage of the detectable domain from the intein domain.

17.   The chimeric protein of claim 1, wherein the intein domain splicing activity results in cleavage of the detectable domain from the intein domain.

18.   The chimeric protein of claim 1, wherein the intein domain has only splicing activity.

19.   The chimeric protein of claim 1, wherein the intein domain has only cleaving activity.

20.   The chimeric protein of claim 1, wherein at least one domain is separated from another domain by a linker.

21.   The chimeric protein of claim 20, wherein the linker is a flexible linker.

22.   The chimeric protein of claim 20, wherein the intein domain is separated from the detectable moiety domain and the enzyme domain by a linker.

23.   An isolated or recombinant nucleic acid encoding a chimeric protein comprising at least three domains, wherein the first domain comprises at least one enzyme domain or a binding protein domain, the second domain comprises at least one intein domain and a third domain comprising a detectable moiety domain, at least one intein domain is positioned between at least one enzyme or binding protein and at least one detectable moiety domain, and the intein domain has at least one splicing or cleavage activity.

24.   An expression cassette comprising an isolated or recombinant nucleic acid encoding a chimeric protein comprising at least three domains, wherein the first domain comprises at least one enzyme domain or a binding protein domain, the second domain comprises at least one intein domain and a third domain comprising a detectable moiety domain, at least one intein domain is positioned between at least one enzyme or binding protein and at least one detectable moiety domain, and the intein domain has at least one splicing or cleavage activity.

25.   A vector comprising an isolated or recombinant nucleic acid encoding a chimeric protein comprising at least three domains, wherein the first domain comprises at least one enzyme domain or a binding protein domain, the second domain comprises at least one intein domain and a third domain comprising a detectable moiety domain, at least one intein domain is positioned between at least one enzyme or binding protein and at least one detectable moiety domain, and the intein domain has at least one splicing activity.

26.   A cell comprising a nucleic acid encoding a chimeric protein comprising at least three domains, wherein the first domain comprises at least one enzyme domain or a binding protein domain, the second domain comprises at least one intein domain and a third domain comprising a detectable moiety domain, at least one intein domain is positioned between at least one enzyme or binding protein and at least one detectable moiety domain, and the intein domain has at least one splicing or cleavage activity.

27.   A non-human transgenic animal comprising a nucleic acid encoding a chimeric protein comprising at least three domains, wherein the first domain comprises at least one enzyme domain or a binding protein domain, the second domain comprises at least one intein domain and a third domain comprising a detectable moiety domain, at least one intein domain is positioned between at least one enzyme or binding protein and at least one detectable moiety domain, and the intein domain has at least one splicing or cleavage activity.

28.   A method for normalizing gene expression comprising the following steps:

(a) providing a nucleic acid encoding a chimeric protein comprising at least three domains, wherein the first domain comprises at least one enzyme domain or a binding protein domain, the second domain comprises at least one intein domain and a third domain comprising a detectable moiety domain, at least one intein domain is positioned between at least one enzyme or binding protein and at least one detectable moiety domain, and the intein domain has at least one splicing or cleavage activity;

(b) expressing the nucleic acid such that the chimeric protein is expressed and the intein domain has at least one splicing activity; and

(c) measuring both the activity of the enzyme and the amount of detectable moiety domain expressed, thereby normalizing gene expression.

29.   The method of claim 28, wherein the detectable moiety domain comprises a detectable peptide or polypeptide.

30.   The method of claim 29, wherein the detectable peptide or polypeptide is a fluorescent peptide or polypeptide.

31.   The method of claim 29, wherein the fluorescent peptide or polypeptide is expressed in a cell and the fluorescence is measured by a FACS.

32.   The method of claim 28, wherein the nucleic acid is expressed *in vivo*.

33.   The method of claim 28, wherein the nucleic acid is expressed *in vitro*.

34.   A high throughput enzymatic screen to measure enzyme activity comprising the following steps:
(a) providing a nucleic acid encoding a chimeric protein comprising at least three domains, wherein the first domain comprises at least one enzyme domain or a binding protein domain, the second domain comprises at least one intein domain and a third domain comprising a detectable moiety domain, at least one intein domain is positioned between at least one enzyme or binding protein and at least one detectable moiety domain, and the intein domain has at least one splicing or cleavage activity;
(b) expressing the nucleic acid *in vivo* or *in vitro*; and
(c) measuring both the activity of the enzyme and the amount of detectable moiety domain expressed.

35.   An isolated or recombinant polypeptide comprising
(a) a polypeptide comprising an amino acid sequence having at least 90% identity to SEQ ID NO:2 over a region of at least about 100 residues, or
(b) a polypeptide encoded by a nucleic acid comprising
(i) a nucleic acid sequence having at least 90% sequence identity to SEQ ID NO:5 over a region of at least about 100 residues; or,
(ii) a nucleic acid that hybridizes under stringent conditions to a nucleic acid comprising a sequence as set forth in SEQ ID NO:1, or a subsequence thereof, wherein the sequence identities are determined by analysis with a sequence comparison algorithm or by a visual inspection and the polypeptide has an intein activity.

36.   The isolated or recombinant polypeptide of claim 35, wherein the intein activity is a cleaving activity or a splicing activity.

37.   The isolated or recombinant polypeptide of claim 35, wherein the polypeptide comprises an amino acid sequence having at least 95% identity to SEQ ID NO:2 over a region of at least about 100 residues.

38.   The isolated or recombinant polypeptide of claim 37, wherein the polypeptide comprises an amino acid sequence having at least 98% identity to SEQ ID NO:2 over a region of at least about 100 residues.

39.   The isolated or recombinant polypeptide of claim 35, wherein the polypeptide comprises an amino acid sequence having at least 90% identity to SEQ ID NO:2 over a region of at least about 200 residues.

40.   The isolated or recombinant polypeptide of claim 35, wherein the polypeptide comprises an amino acid sequence having at least 90% identity to SEQ ID NO:2 over a region of at least about 300 residues.

41.   The isolated or recombinant polypeptide of claim 35, wherein the polypeptide comprises an amino acid sequence having at least 90% identity to SEQ ID NO:2 over a region of at least about 400 residues.

42.   The isolated or recombinant polypeptide of claim 35, wherein the polypeptide comprises an amino acid sequence having at least 90% identity to SEQ ID NO:2 over a region of at least about 500 residues.

43.   The isolated or recombinant polypeptide of claim 35, wherein the polypeptide comprises an amino acid sequence having at least 90% identity to SEQ ID NO:2 over a region of at least about 600 residues.

44.   The isolated or recombinant polypeptide of claim 35, wherein the polypeptide comprises an amino acid sequence having at least 90% identity to SEQ ID NO:2 over a region of at least about 300 residues.

45.   The isolated or recombinant polypeptide of claim 35, wherein the an amino acid sequence as set forth in SEQ ID NO:2.

46.   An isolated or recombinant nucleic acid comprising a nucleic acid sequence having at least 90% sequence identity to SEQ ID NO:1 over a region of at least about 100 residues, wherein the nucleic acid encodes at least one polypeptide having an intein activity, and the sequence identities are determined by analysis with a sequence comparison algorithm or by a visual inspection.

47.   The isolated or recombinant nucleic acid of claim 46, wherein the nucleic acid sequence has at least 95% sequence identity to SEQ ID NO:1 over a region of at least about 100 residues.

48.   The isolated or recombinant nucleic acid of claim 47, wherein the nucleic acid sequence has at least 98% sequence identity to SEQ ID NO:1 over a region of at least about 100 residues.

49.   The isolated or recombinant nucleic acid of claim 48, wherein the nucleic acid sequence has a sequence as set forth in SEQ ID NO:1.

50.   The isolated or recombinant nucleic acid of claim 46, wherein the nucleic acid sequence has at least 90% sequence identity to SEQ ID NO:1 over a region of at least about 200 residues.

51.   The isolated or recombinant nucleic acid of claim 50, wherein the nucleic acid sequence has at least 90% sequence identity to SEQ ID NO:1 over a region of at least about 300 residues.

52.   The isolated or recombinant nucleic acid of claim 51, wherein the nucleic acid sequence has at least 90% sequence identity to SEQ ID NO:1 over a region of at least about 400 residues.

53.   The isolated or recombinant nucleic acid of claim 52, wherein the nucleic acid sequence has at least 90% sequence identity to SEQ ID NO:1 over a region of at least about 500, 600, 700, 800, 900 or 1000 residues.

54.   The isolated or recombinant nucleic acid of claim 46, wherein the sequence comparison algorithm is a BLAST version 2.2.2 algorithm where a filtering setting is set to blastall -p blastp -d "nr pataa" -F F, and all other options are set to default.

55.   An isolated or recombinant nucleic acid, wherein the nucleic acid comprises a sequence that hybridizes under stringent conditions to a nucleic acid comprising a sequence as set forth in SEQ ID NO:1.

56.   The isolated or recombinant nucleic acid of claim 55, wherein the nucleic acid is at least about 100 residues in length.

57.   The isolated or recombinant nucleic acid of claim 56, wherein the nucleic acid is at least about 200 residues in length.

58.   The isolated or recombinant nucleic acid of claim 57, wherein the nucleic acid is at least about 200, 300, 400 residues in length.

59.   The isolated or recombinant nucleic acid of claim 58, wherein the nucleic acid is at least about 500, 600, 700, 800, 900, 1000 residues in length or the full length of the gene or transcript.

60.   The isolated or recombinant nucleic acid of claim 55, wherein the stringent conditions include a wash step comprising a wash in 0.2X SSC at a temperature of about 65oC for about 15 minutes.

61.   A nucleic acid probe for identifying a nucleic acid encoding a polypeptide with an intein activity, wherein the probe comprises at least 10 consecutive bases of a sequence comprising a sequence as set forth in SEQ ID NO:1.

62. An amplification primer sequence pair for amplifying a nucleic acid encoding a polypeptide having an intein activity, wherein the primer pair is capable of amplifying a nucleic acid comprising a sequence as set forth in SEQ ID NO:1.

63. A method of amplifying a nucleic acid encoding a polypeptide having an intein activity comprising amplification of a template nucleic acid with an amplification primer sequence pair capable of amplifying a nucleic acid sequence as set forth in SEQ ID NO:1, or a subsequence thereof.

64. An expression cassette comprising a nucleic acid sequence having at least 90% sequence identity to SEQ ID NO:1 over a region of at least about 100 residues, wherein the nucleic acid encodes at least one polypeptide having an intein activity, and the sequence identities are determined by analysis with a sequence comparison algorithm or by a visual inspection.

65. A vector comprising a nucleic acid sequence having at least 90% sequence identity to SEQ ID NO:1 over a region of at least about 100 residues, wherein the nucleic acid encodes at least one polypeptide having an intein activity, and the sequence identities are determined by analysis with a sequence comparison algorithm or by a visual inspection.

66. A cloning vehicle comprising a vector as set forth in claim 65, wherein the cloning vehicle comprises a viral vector, a plasmid, a phage, a phagemid, a cosmid, a fosmid, a bacteriophage or an artificial chromosome.

67. A transformed cell comprising the expression cassette of claim 64 or the vector of claim 65.

68. The transformed cell of claim 67, wherein the cell is a bacterial cell, a mammalian cell, a fungal cell, a yeast cell, an insect cell or a plant cell.

69. A transgenic non-human animal comprising a nucleic acid sequence having at least 90% sequence identity to SEQ ID NO:1 over a region of at least about 100 residues, wherein the nucleic acid encodes at least one polypeptide having an

intein activity, and the sequence identities are determined by analysis with a sequence comparison algorithm or by a visual inspection.

70.   A transgenic plant comprising a nucleic acid sequence having at least 90% sequence identity to SEQ ID NO:1 over a region of at least about 100 residues, wherein the nucleic acid encodes at least one polypeptide having an intein activity, and the sequence identities are determined by analysis with a sequence comparison algorithm or by a visual inspection.

71.   A method of inhibiting the translation of an intein message in a cell comprising administering to the cell or expressing in the cell an antisense oligonucleotide comprising a nucleic acid sequence complementary to or capable of hybridizing under stringent conditions to a nucleic acid having at least 90% sequence identity to SEQ ID NO:1 over a region of at least about 100 residues, wherein the nucleic acid encodes at least one polypeptide having an intein activity, and the sequence identities are determined by analysis with a sequence comparison algorithm or by a visual inspection.

72.   An isolated or recombinant antibody that specifically binds to a polypeptide as set forth in claim 35 or to a polypeptide encoded by a nucleic acid as set forth in claim 46 or claim 55.

73.   An isolated or recombinant polypeptide comprising a polypeptide comprising an amino acid sequence having at least 90% identity to SEQ ID NO:2 over a region of at least about 100 residues with the proviso that it is not associated with any sequence to which it is naturally associated on its amino terminal end, on its carboxy terminal end, or on both ends of the polypeptide.

74.   The isolated or recombinant polypeptide of claim 73, wherein the polypeptide is not associated with an *Aquifex* sequence.

75.   An isolated or recombinant nucleic acid comprising a sequence having at least 90% identity to SEQ ID NO:1 over a region of at least about 100 residues with the proviso that it is not associated with any sequence to which it is naturally associated on its 3' end, on its 5' end, or on both ends of the sequence.

76.   The isolated or recombinant nucleic acid of claim 75, wherein the nucleic acid is not associated with an *Aquifex* sequence.

77.   The isolated or recombinant nucleic acid of claim 75, wherein the nucleic acid is not associated with an *Aquifex* sequence on its 3' end.

78.   The isolated or recombinant nucleic acid of claim 75, wherein the nucleic acid is not associated with an *Aquifex* sequence on its 5' end.

INTEIN SEQUENCE LISTING.txt

SEQUENCE LISTING

<110>

<120>

<130>

<160> 2

<170> FastSEQ for Windows Version 4.0

<210> 1
<211> 2091
<212> DNA
<213> Aquifex aeolicus

<400> 1
```
atggaaaga cagaaaaaaa tgagcttgtc agaaaactca ttttcaaccc tcaaggagac     60
agggaggcga gcaaaaggaa gataataaag ggaaacccga caaacatatt tgaacttaac    120
gagataaagt attcctgggc ttttgacctt tacaagttaa tgggctttac aaacttctgg    180
atacccgaag agatacagat gcttgaagac aggaaacagt acagaaccgt tctatcagac    240
tacgaaaaga gggcatacga actcgtcctt tccttcctca tagcccttga ctcctttcaa    300
gtggacatgc ttaaagagtt cggaaggatg ataaccgccc ccgaagtaga aatggccata    360
acagctcagg aatttcagga atccgtccac gcgtactctt accagttcat actcgagtct    420
gtagttgatc cggttaaagc ggacgagatt tacaactact ggcgggagga tgaaagactt    480
ctggaaagga ataaagtaat agcagacgtg tacaacgaat tcattagaaa acccaacgaa    540
gaaaactta taaaggcaac aatagggaac tacatactcg agagcctgta cttttactct    600
ggatttgcct tcttctacac actgggaaga cagggcaaaa tgagaaacac tgtacagcaa    660
atcaaatata tcaacaggga tgagctctgc ttcattgagg aacggaggt tttgacgaag    720
aggggtttcg ttgatttcag ggagctgagg gaagacgatc ttgtagctca gtacgatata    780
gaaacaggg aaatttcctg gacaaaacct tacgcctacg ttgaaaggga ttacgagggt    840
tctatgtaca gattaaaaca tcctaaaagc aactgggaag tagtagctac tgaagggcac    900
gagttcatag taaggaacct gaaaacagga aaggagagaa aggaaccgat agaaaaggta    960
aaactacatc cctactctgc aattcccgtt gcgggaaggt acacgggaga agtggaagag   1020
tacgacctct gggaactcgt aagcggaaaa ggtataactc ttaaaaccga gagtgctgtg   1080
aagaataagt taacaccgat agaaaaactc ctgatagttc ttcaggcgga cgggacaata   1140
gacagtaaga gaaatggaaa gttcacaggc ttccaacaat taaagttctt cttctcaaag   1200
tatagaaaga ttaacgagtt tgaaaaaata ctccaatgaat gtgcacctta cggaattaaa   1260
tggaaaaagt acgagcgcca agacggaatt gcttacacag tttactatcc gaatgacctt   1320
ccgataaagc ctactaagtt ctttgacgaa tgggtgagac ttgatgagat aacggaagaa   1380
tggataaggg aatttgtgga agaactcgtc aagtgggacg gacacattcc gaaagacagg   1440
aataaaaaga aggtttatta ttactccaca aaagaaaaaa gaaacaagga ctttgtgcag   1500
gcactttgtg ctctgggagg tatgagaact gttgtcagta gagagagaaa tccgaaggcg   1560
aaaaaccccg tttacaggat atggatttac ctagaggacg actacataaa taccaaaaca   1620
atggtgaagg aagagttcta ctacaaaggt aaggtgtact gcgtgagcgt tcccaaaggg   1680
aacatagttg tgagtacaca agacagcgtt tgtattgcg gcaactgcca cgttacgctc   1740
ttcaggaaca taataaaacac attcaggaaa agaaatcccg aattatttac gcctgagata   1800
gaaaagtgga tagtgggata cttcaagtac gcgtgaaacg aagaaatcaa atggggggcag   1860
tatgttaccc agaaccagat actcggtatt aacgacgtct tgatagagag gtatataaag   1920
tatctcggaa acctgaggat tactcagatc ggctttgatc cgatatatcc agaggttaca   1980
gaaaaccccct taaagtggat agacgagttt agaaagataa acaacactaa aacggacttc   2040
ttccaggcaa agcctcagac ctactccaaa gccaacgaac tcaagtggta a             2091
```

<210> 2
<211> 696
<212> PRT
<213> Aquifex aeolicus

<400> 2
```
Met Glu Lys Thr Glu Lys Asn Glu Leu Val Arg Lys Leu Ile Phe Asn
1               5                   10                  15
```

INTEIN SEQUENCE LISTING.txt

```
Pro Gln Gly Asp Arg Glu Ala Ser Lys Arg Lys Ile Ile Lys Gly Asn
        20                  25                  30
Pro Thr Asn Ile Phe Glu Leu Asn Glu Leu Lys Tyr Ser Trp Ala Phe
        35                  40                  45
Asp Leu Tyr Lys Leu Met Gly Phe Thr Asn Phe Trp Ile Pro Glu Glu
50                  55                  60
Ile Gln Met Leu Glu Asp Arg Lys Gln Tyr Glu Thr Val Leu Ser Asp
65                  70                  75                  80
Tyr Glu Lys Arg Ala Tyr Glu Leu Val Leu Ser Phe Leu Ile Ala Leu
            85                  90                  95
Asp Ser Phe Gln Val Asp Met Leu Lys Glu Phe Gly Arg Met Ile Thr
            100                 105                 110
Ala Pro Glu Val Glu Met Ala Ile Thr Ala Gln Glu Phe Gln Glu Ser
        115                 120                 125
Val His Ala Tyr Ser Tyr Gln Phe Ile Leu Glu Ser Val Val Asp Pro
    130                 135                 140
Val Lys Ala Asp Glu Ile Tyr Asn Tyr Trp Arg Glu Asp Glu Arg Leu
145                 150                 155                 160
Leu Glu Arg Asn Lys Val Ile Ala Glu Leu Tyr Asn Glu Phe Ile Arg
                165                 170                 175
Lys Pro Asn Glu Glu Asn Pro Ile Lys Ala Thr Ile Gly Asn Tyr Ile
            180                 185                 190
Leu Glu Ser Leu Tyr Phe Tyr Ser Gly Phe Ala Phe Phe Tyr Thr Leu
        195                 200                 205
Gly Arg Gln Gly Lys Met Arg Asn Thr Val Gln Gln Ile Lys Tyr Ile
    210                 215                 220
Asn Arg Asp Glu Leu Cys Phe Ile Glu Gly Thr Glu Val Leu Thr Lys
225                 230                 235                 240
Arg Gly Phe Val Asp Phe Arg Glu Leu Arg Glu Asp Asp Leu Val Ala
                245                 250                 255
Gln Tyr Asp Ile Glu Thr Gly Glu Ile Ser Trp Thr Lys Pro Tyr Ala
            260                 265                 270
Tyr Val Glu Arg Asp Tyr Glu Gly Ser Met Tyr Arg Leu Lys His Pro
        275                 280                 285
Lys Ser Asn Trp Glu Val Val Ala Thr Glu Gly His Glu Phe Ile Val
    290                 295                 300
Arg Asn Leu Lys Thr Gly Lys Glu Arg Lys Glu Pro Ile Glu Lys Val
305                 310                 315                 320
Lys Leu His Pro Tyr Ser Ala Ile Pro Val Ala Gly Arg Tyr Thr Gly
            325                 330                 335
Glu Val Glu Glu Tyr Asp Leu Trp Glu Leu Val Ser Gly Lys Gly Ile
        340                 345                 350
Thr Glu Lys Thr Arg Ser Ala Val Lys Asn Lys Leu Thr Pro Ile Glu
    355                 360                 365
Lys Leu Leu Ile Val Leu Gln Ala Asp Gly Thr Ile Asp Ser Lys Arg
370                 375                 380
Asn Gly Lys Phe Thr Gly Phe Gln Gln Leu Lys Phe Phe Phe Ser Lys
385                 390                 395                 400
Tyr Arg Lys Ile Asn Glu Phe Glu Lys Ile Leu Asn Glu Cys Ala Pro
                405                 410                 415
Tyr Gly Ile Lys Trp Lys Lys Tyr Glu Arg Gln Asp Gly Ile Ala Tyr
            420                 425                 430
Thr Val Tyr Tyr Pro Asn Asp Leu Pro Ile Lys Pro Thr Lys Phe Phe
        435                 440                 445
Asp Glu Trp Val Arg Leu Asp Glu Ile Thr Glu Glu Trp Ile Arg Glu
    450                 455                 460
Phe Val Glu Glu Leu Val Lys Trp Asp Gly His Ile Pro Lys Asp Arg
465                 470                 475                 480
Asn Lys Lys Lys Val Tyr Tyr Tyr Ser Thr Lys Glu Lys Arg Asn Lys
                485                 490                 495
Asp Phe Val Gln Ala Leu Cys Ala Leu Gly Gly Met Arg Thr Val Val
            500                 505                 510
Ser Arg Glu Arg Asn Pro Lys Ala Lys Asn Pro Val Tyr Arg Ile Trp
```

```
                                           INTEIN SEQUENCE LISTING.txt
            515                            520                    525
Ile Tyr Leu Glu Asp Asp Tyr Ile Asn Thr Gln Thr Met Val Lys Glu
        530                         535                    540
Glu Phe Tyr Tyr Lys Gly Lys Val Tyr Cys Val Ser Val Pro Lys Gly
545                         550                    555                    560
Asn Ile Val Val Arg Tyr Lys Asp Ser Val Cys Ile Ala Gly Asn Cys
                    565                         570                    575
His Val Thr Leu Phe Arg Asn Ile Ile Asn Thr Leu Arg Lys Glu Asn
                580                         585                    590
Pro Glu Leu Phe Thr Pro Glu Ile Glu Lys Trp Ile Val Glu Tyr Phe
        595                         600                    605
Lys Tyr Ala Val Asn Glu Glu Ile Lys Trp Gly Gln Tyr Val Thr Gln
        610                         615                    620
Asn Gln Ile Leu Gly Ile Asn Asp Val Leu Ile Glu Arg Tyr Ile Lys
625                         630                    635                    640
Tyr Leu Gly Asn Leu Arg Ile Thr Gln Ile Gly Phe Asp Pro Ile Tyr
                    645                         650                    655
Pro Glu Val Thr Glu Asn Pro Leu Lys Trp Ile Asp Glu Phe Arg Lys
                660                         665                    670
Ile Asn Asn Thr Lys Thr Asp Phe Phe Gln Ala Lys Pro Gln Thr Tyr
        675                         680                    685
Ser Lys Ala Asn Glu Leu Lys Trp
        690                         695
```